# Actes des 29es journées francophones d'Ingénierie des Connaissances

Sylvie Ranwez

APIA & CNIA & IC & JFPDA & RJCIA

IA pour l'éducation IA & Santé TAL & IA Ethique & IA
Robotique & IA France@IJCAI2018

# PFIA 2018

11e Plate-forme
Intelligence Artificielle

2 au 6 juillet 2018 - Nancy

Campus Sciences - Université de Lorraine
Vandœuvre-lès-Nancy

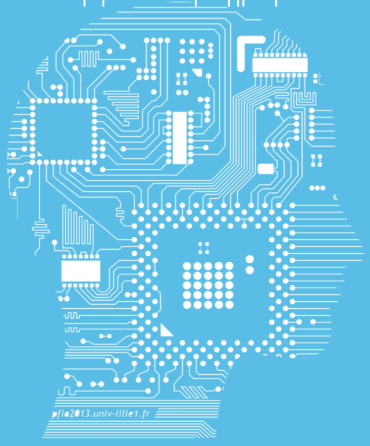
# IC 2018

Actes

des

## 29e Journées Francophones d'Ingénierie des Connaissances

Présidente du comité de programme
Sylvie Ranwez



PFIA2018.LORIA.FR

## Comité de programme

## Présidente

- Sylvie Ranwez, LGI2P, IMT Mines Alès

## Comité de programme

- Marie-Hélène Abel, HEUDIASYC, Université de Technologie de Compiègne
- Xavier Aimé, LIMICS, Université Paris 13
- Yamine Ait Ameur, IRIT, INP Toulouse
- Florence Amardeilh, Créatrice d'entreprise en Ingénierie des Connaissances
- Bruno Bachimont, COSTECH, Université de Technologie de Compiègne
- Jean-Paul Barthès, HEUDIASYC, Université de Technologie de Compiègne
- Aurélien Bénel, ICD, Université de Technologie de Troyes
- Nacéra Bennacer, LRI, Centrale Supélec campus de Gif-sur-Yvette
- Pierre Bourhis, LIFL, CNRS
- Bertrand Braunschweig, Inria Saclay-Île-de-France
- Nathalie Bricon-Souf, IRIT, Université Paul Sabatier Toulouse 3
- Sandra Bringay, LIRMM, Université Paul-Valéry Montpellier 3
- Patrice Buche, IATE, INRA Montpellier
- Davide Buscaldi, LIPN, Université Paris 13
- Elena Cabrio, I3S, INRIA Sophia Antipolis - Méditerranée
- Sylvie Calabretto, LIRIS, INSA de Lyon
- Gaoussou Camara, Université Alioune Diop de Bambey
- Pierre-Antoine Champin, LIRIS, Université Claude Bernard Lyon 1
- Jean-Pierre Chanet, TSCF, Irstea de Clermont Ferrand
- Jean Charlet, LIMICS, APHP INSERM Paris
- Olivier Corby, Université Côte d'Azur, Inria
- Amélie Cordier, Hoomano, Université de Lyon
- Mathieu D'Aquin, Insight Centre for Data Analytics, NUI Galway, Ireland
- Jérôme David, LIG, INRIA Grenoble
- Sylvie Despres, LIMICS, Université Paris 13
- Rim Djedidi, LIMICS, Université Paris 13
- Jean-Pierre Evain, EBU Suisse
- Gilles Falquet, CUI, Université de Genève
- Catherine Faron-Zucker, I3S, Université Nice Sophia Antipolis
- Cécile Favre, ERIC, Université Lumière Lyon 2
- Béatrice Fuchs, LIRIS, IAE - université Lyon 3
- Frédéric Fürst, MIS, Université de Picardie Jules Verne
- Jean-Gabriel Ganascia, LIP6, Université Pierre et Marie Curie
- Serge Garlatti, IMT Atlantique
- Alain Giboin, I3S, INRIA Sophia Antipolis - Méditerranée
- Nathalie Guin, LIRIS, Université Claude Bernard Lyon 1
- Ollivier Haemmerlé, IRIT, Université Toulouse le Mirail
- Sébastien Harispe, LGI2P, IMT Mines Alès
- Mounira Harzallah, LS2N, Université de Nantes
- Nathalie Hernandez, IRIT, Université Toulouse le Mirail
- Liliana Ibanescu, MIA, INRA AgroParistech Paris

- Sébastien Iksal, LIUM, IUT Le Mans
- Antoine Isaac, Europeana & VU University Amsterdam
- Clement Jonquet, LIRMM, Université de Montpellier
- Mouna Kamel, IRIT, Université de Perpignan Via Domitia
- Gilles Kassel, MIS, Université de Picardie Jules Verne
- Pascale Kuntz, LS2N, Université de Nantes
- Florence Le Ber, ICUBE, Université de Strasbourg / ENGEES
- Michel Leclère, LIRMM, Université de Montpellier
- Maxime Lefrançois, laboratoire Hubert Curien, IMT Mines Saint-Étienne
- Alain Léger, Orange Labs, France Telecom Rennes
- Dominique Lenne, HEUDIASYC, Université de Technologie de Compiègne
- Moussa Lo, Université Gaston Berger de Saint Louis
- Cédric Lopez, Emvista, Montpellier
- Nada Matta, ICD, Université de Technologie de Troyes
- Pascal Molli, LS2N, Université de Nantes
- Alexandre Monnin, Origens Medialab et ESC Clermont
- Isabelle Mougenot, UMR Espace dev, Université de Montpellier
- Fleur Mougin, ERIAS/INSERM BPH U1219, Université de Bordeaux
- Amedeo Napoli, LORIA, CNRS Nancy
- Emmanuel Nauer, LORIA, Université de Metz
- Jérôme Nobécourt, LIMICS, Université Paris 13
- Nathalie Pernelle, LRI, Université Paris Sud
- Yannick Prié, LS2N, Université de Nantes
- Cédric Pruski, Luxembourg Institute of Science and Technology
- Chantal Reynaud, LRI, Université Paris Sud
- Catherine Roussey, TSCF, Irstea
- Fatiha Saïs, LRI, Université Paris Sud
- Pascal Salembier, ICD, Université de Technologie de Troyes
- Karim Sehaba, LIRIS, Université Lumière Lyon 2
- Hassina Seridi-Bouchelaghem, LabGED, Badji Mokhtar University
- Andrea Tettamanzi, I3S, Université Nice Sophia Antipolis
- Konstantin Todorov, LIRMM, Université de Montpellier
- Francky Trichet, LS2N, Université de Nantes
- Raphaël Troncy, Data Science, EURECOM
- Serena Villata, I3S, CNRS
- Amel Yessad, LIP6, Université Pierre et Marie Curie
- Haïfa Zargayouna, LIPN, Université Paris 13
- Pierre Zweigenbaum, LIMSI, CNRS, Université Paris-Saclay, Orsay

# Avant-propos

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

L'importance grandissante du numérique dans la plupart de nos activités humaines conduit au développement de nombreuses applications dites intelligentes et de services qui visent à faciliter nos prises de décision, nous conseillent dans nos actions, nous instruisent, nous protègent ou nous divertissent... Or ces applications, fondées sur des technologies issues de l'Intelligence Artificielle, nécessitent généralement une représentation de la connaissance d'un domaine propre à supporter des raisonnements automatisés, des inférences, des processus d'apprentissage. Hébergée par la plateforme PFIA 2018 (à Nancy du 2 au 6 juillet), la 29$^{e}$ édition de la conférence IC souhaite mettre en lumière toutes les approches liées à l'ingénierie de connaissance au sein d'applications de l'intelligence artificielle que ce soit sur le Web, dans les entreprises ou embarquées sur différents objets connectés. Ces actes regroupent les visions croisées de chercheurs académiques et d'industriels sur différents aspects de l'Ingénierie des Connaissances, tant dans les fondements théoriques que dans leur mise en œuvre.

L'édition 2018 regroupe 12 articles longs et 5 articles courts sélectionnés parmi 28 soumissions (soit un taux d'acceptation de 60%), parmi lesquels on retrouve de nombreuses contributions liées à la construction d'ontologies et à leur intégration dans différents systèmes d'information liés à l'évolution de nos sociétés, de nos modes de vies et leur impact sur notre environnement. Les sessions proposées concernent les thèmes suivants :

- Représentation des connaissances sur la Terre : une contribution au développement durable ?
- Extraction de connaissances pour aider la décision.
- Différentes méthodologies pour la conception d'ontologies.
- Construction d'ontologies : des fondements théoriques à la mise en œuvre.
- Construction collaborative d'ontologies qui intègre l'utilisateur et son contexte

Les communications affichées (8 posters) sont regroupées en fin de document. Elles ont été sélectionnées pour susciter des discussions autour d'approches ou de travaux non encore aboutis, mais de grand intérêt tant dans l'originalité de la démarche évoquée que dans la pertinence de l'application qui en découle.

Mes remerciements les plus sincères vont à toutes celles et tous ceux sans qui ces journées ne pourraient avoir lieu : Nicola Guarino qui a accepté de partager avec la communauté francophone une longue expérience dans la conception et l'usage des ontologies, les auteurs qui ont contribué à cet ouvrage, le comité de programme dont la qualité des relectures a été fortement appréciée de tous.

J'adresse un grand merci aux intervenants de la table ronde "Ingénierie des connaissances : composante indispensable de l'Intelligence Artificielle ?" qui ont accepté avec une extrême gentillesse de partager avec nous leur vision de l'intelligence artificielle et de la place qu'occupe l'IC dans ses applications : Céline Rouveirol, Alain Berger, Jean Charlet et Laurent Pierre.

Merci à l'AFIA qui supporte ces journées (prix du meilleur papier) et nous héberge cette année *via* la plateforme PFIA, au comité d'organisation et à ses deux présidents Armelle et Davy, au chef d'orchestre, Yves demazeau, et à l'ensemble des présidents des autres conférences et journées organisées sur la plateforme et qui, par nos échanges, ont permis une belle coordination. Enfin, merci à l'ensemble des organismes qui soutiennent cet évènement.

Merci à toutes celles et tous ceux qui, dans l'ombre, ont été de bons conseils pour prendre certaines décisions…

Que ces journées et la lecture de ces actes soient le ferment d'idées nouvelles et de discussions riches pour contribuer, ensemble, au développement de l'Ingénierie des Connaissances.

Sylvie Ranwez
Présidente du comité de programme

# Sommaire

# 25 Years of Applied Ontology and Ontological Analysis: an Interdisciplinary Endeavour

Nicola Guarino - Directeur de Recherche (CNR, Italie)

**Abstract:** Applied Ontology is an emerging discipline –born about 25 years ago– that builds on philosophy, cognitive science, linguistics and logic with the purpose of understanding, clarifying, making explicit and communicating people's assumptions about the nature and structure of the world. This orientation towards helping people understanding each other distinguishes applied ontology from philosophical ontology, and motivates its unavoidable interdisciplinary nature. In this talk I will briefly review the problems that applied ontology can address, the conceptual tools at the basis of formal ontological analysis, and the future application perspectives.

**Biography:** Nicola Guarino, research director at the Italian National Research Council (CNR), works at the nation-wide Institute of Cognitive Sciences and Technologies (ISTC-CNR), leading the Laboratory for Applied Ontology (LOA) located in Trento. A graduate in electronic engineering at Padua university in 1978, since 1991 has been playing a leading role in the ontology field, developing a strongly interdisciplinary approach that combines together Computer Science, Philosophy, and Linguistics. His impact is testified by a long list of widely cited papers and many keynote talks and tutorials in major conferences involving different communities. Among the most well known results of his lab, the OntoClean methodology and the DOLCE foundational ontology. On the theoretical side, current research interest include the formal ontology of relationships and events, while on the application side his research is focusing on service science, socio-technical systems, and e-government. He is founder and former editor-in-chief (with Mark Musen of Stanford University) of Applied Ontology, founder and past president of the International Association for Ontology and its Applications, former general chair of the international conference on Formal Ontology in Information Systems (FOIS), editorial board member of Journal of Data Semantics, and editor of the IOS Press book series Frontiers in AI and Applications. He is also fellow of the European Coordinating Committee for Artificial Intelligence (ECCAI).

# Représentation des connaissances sur la Terre : une contribution au développement durable ?

Les applications du numérique dans différents secteurs de nos activités humaines impactent nos modes vie, de production, de consommation et ont fatalement des impacts sur notre environnement. Ces répercussions peuvent être négatives et il convient alors de les limiter, mais elles peuvent également être positives : en couplant des données satellitaires, la connaissance des domaines concernés et des données de terrain, il est possible de combattre plus efficacement les incendies, de mieux gérer des territoires agricoles, de favoriser une irrigation responsable et contrôlée. Les défis dans ces secteurs sont nombreux du point de vue de la recherche. Les données sont hétérogènes, peuvent être variables dans le temps, et leur traitement nécessite des adaptations qui sont discutées dans cette session.

# Ontologie pour l'intégration de données d'observation de la Terre et contextuelles basée sur les relations topologiques

Helbert Arenas, Nathalie Aussenac-Gilles, Catherine Comparot, Cassia Trojahn

IRIT, CNRS, UNIVERSITÉ TOULOUSE 2 ET UNIVERSITÉ DE TOULOUSE France
{prenom.nom}@irit.fr

**Résumé** : Nous proposons une ontologie pour l'intégration de données d'observation de la Terre et de donnés contextuelles, à l'aide de relations topologiques spatiales et temporelles. Cette ontologie spécialise des standards, notamment SOSA, GeoSPARQL et OWL-Time. La base de connaissance décrite par cette ontologie est alimentée grâce à un processus qui sélectionne, transforme et intègre de données géospatiales hétérogènes (méta-données d'image satellite, données météorologiques, unités administratives, couverture terrestre, etc.). Ce processus s'appuie sur le tuilage des images pour traiter les données ayant une composante spatiale fixe, alors que les relations temporelles sont calculées à la volée à partir d'une topologie temporelle. Nous validons cette approche par un cas d'étude exploitant des méta-données d'image satellite Sentinel.
**Mots-clés** : Ontologies, intégration de données, données spatiales et temporelles, observation de la Terre, SOSA.

## 1 Introduction

L'observation de la Terre offre une valeur ajoutée à une grande diversité de domaines. Récemment l'Agence Spatiale Européenne (ESA) a lancé le programme Sentinel avec deux types de satellites, Sentinel-1 and Sentinel-2, qui transmettent des images de haute qualité (entre 8 à 10 To de données quotidiennement). Ces images de la Terre sont captées selon différentes technologies et libres d'accès. Cette disponibilité des données ouvre de nombreuses perspectives économiques grâce à de nouvelles applications dans des domaines aussi variés que l'agriculture, l'environnement, l'urbanisme, l'océanographie ou encore la climatologie. Ces applications métier ont néanmoins un besoin de coupler les images avec des données sur les zones observées. Ces données sont accessibles à partir de différentes sources dans des formats hétérogènes et des temporalités différentes : elles peuvent être statiques, comme les données sur la couverture terrestre, ou dynamiques, comme les observations météo. Elles peuvent être utiles par exemple pour indiquer qu'une image contient une région touchée par un phénomène tel qu'un tremblement de terre ou une canicule, et ensuite pour décider des actions à mener dans cette zone ou conduire à des analyses à plus long terme. Plus encore, en exploitant les caractéristiques spatio-temporelles d'un phénomène (son empreinte spatiale et sa date), il devient possible de savoir si une entité localisée dans l'empreinte de l'image (e.g. une ville) a subi le même phénomène.

Les images satellites étant décrites par des méta-données, le problème revient à intégrer à ces méta-données des données provenant de sources variées et très hétérogènes (format, représentation). L'apport des technologies sémantiques pour faciliter cette tâche a été démontré dans des travaux antérieurs (Reitsma & Albrecht, 2005; Sukhobok *et al.*, 2017), en particulier grâce à l'utilisation d'ontologies comme représentations formelles des connaissances d'un domaine donné. Ainsi, dans la continuité des travaux sur l'accès et l'intégration de données via les ontologies (i.e.OBDA pour "Ontology-based Data Access" et OBDI pour "Ontology-based Data Integration") (Lenzerini, 2011; Lefrançois *et al.*, 2017; Console & Lenzerini, 2014), nous avons conçu un vocabulaire sémantique permettant de représenter les données des différentes sources envisagées et d'y accéder de façon homogène. Notre approche a consisté à construire une ontologie modulaire permettant de répondre aux besoins

d'exploitation propres à chaque source de données. Cette approche permet de réduire partiellement la complexité et l'hétérogénéité des données, et ainsi de faciliter le peuplement de l'ontologie. Une caractéristique commune aux observations de la Terre est qu'elles peuvent être liées via leurs propriétés topologiques spatiales et temporelles. Un préalable au peuplement de l'ontologie (i.e. la production d'entrepôts RDF), a été de convertir les différents ensembles de données géo-spatiales fournis dans des formats hétérogènes (shapefile, KML, CSV, GeoJSON, TIFF), en un format commun, JSON, voire d'assurer, le cas échéant, la compatibilité de leurs propriétés et relations spatiales et temporelles, puis de les stocker dans une base de données NoSQL MongoDB.

Nous présentons dans cet article le vocabulaire que nous avons défini pour assurer la description sémantique homogène des différents types de données sous forme d'entités ayant des propriétés spatiales et temporelles. Une partie des données géospatiales à intégrer aux méta-données d'image sont des données contextuelles mesurées à la surface de la Terre, que nous traitons comme des données de capteurs. Ce vocabulaire spécialise ainsi des vocabulaires connus du LOD, dont SOSA [1] pour les données de capteurs, GeoSPARQL (Kolas *et al.*, 2013) pour gérer les relations et coordonnées spatiales, et OWL Time [2] pour traiter la dimension temporelle. De fait, nous nous sommes appuyés sur une topologie des entités, en distinguant notamment les données dynamiques, dont la validité est fournie par la composante temporelle, et les données statiques, pour lesquelles la composante spatiale joue un rôle plus primordial.

Nous montrons, à travers un cas d'étude, comment cette ontologie permet de prendre en compte les spécificités de différentes sources en termes de vocabulaire, périodicité et volume, et comment elle peut améliorer l'accès aux données. A titre d'exemple, nous avons généré des entrepôts RDF qui mettent en relation des informations sur la couverture terrestre (% de forêts, d'eau, etc.) d'une part, et des relevés de stations météorologiques (température, humidité, etc.) d'autre part, avec les images satellites. Ce travail est mené dans le cadre du projet SparkinData visant à construire une plate-forme cloud de données d'observations de la Terre. Le cas d'étude s'appuie sur l'annotation sémantique de méta-données d'images brutes fournies par le CNES (Centre National d'Etudes Spatiales).

Le reste de l'article est organisé ainsi : la partie 2 expose des travaux liés ; en partie 3, nous présentons le modèle d'intégration ; son exploitation pour représenter en RDF les diverses sources d'informations envisagées est l'objet de la partie 4 ; la partie 5 décrit différentes stratégies mises en oeuvre pour intégrer les données des entrepôts RDF via les relations topologiques. Enfin, nous concluons et présentons des perspectives à ce travail en partie 6.

## 2 Travaux liés

### 2.1 Modèle sémantique pour l'imagerie satellitaire

Le projet européen TELIOS (Virtual Observatory Infrastructure for Earth Observation Data) [3] a été pionner dans l'utilisation de représentations sémantiques pour décrire des images de satellites et faciliter l'accès à des données d'observation de la Terre par des agences telles que la NASA ou l'ESA. Le cadre applicatif de ce projet est la gestion des incendies en Grèce (The TELEIOS Team, 2012). Les images sont classifiées et représentées sous forme vectorielle, et à chaque vecteur est associée une estimation des risques d'incendie. Les vecteurs sont liés à des sources de données externes par le biais d'ontologies (TELEIOS, 2016). Ces ontologies exploitent notamment le standard stRDF qui étend RDF pour les données spatiales et temporelles ; seule la composante spatiale est exploitée dans les analyses réalisées.

Le travail de Espinoza-Molina & Datcu (2013) et Espinoza-Molina *et al.* (2015) a été influencé par TELEIOS. Leur système permet d'ajouter des annotations sémantiques à des images produites par des radars à synthèse d'ouverture (SAR) selon le processus suivant :

1. `https://www.w3.org/2015/spatial/wiki/SOSA\_Ontology`
2. `https://www.w3.org/TR/owl-time/(10/2017)`
3. http://www.earthobservatory.eu Accès : 2016-12-01

1) extraction des informations pertinentes à partir des fichiers de métadonnées des images (fichiers XML) ; 2) division de l'image en *tuiles* de tailles différentes et organisation de ces tuiles sous une forme pyramidale ; 3) extraction puis transformation sous forme de vecteurs des caractéristiques de chaque tuile. Ce système est capable de trouver de l'information en utilisant des requêtes portant sur les méta-données, la sémantique et le contenu spatial des images. Cette approche associe aux images d'observation de la Terre, des données provenant des sources externes comme Corine Land Cover, Urban Atlas, Geonames et LinkedGeoData. Elle s'appuie sur un modèle ontologique, SAR, qui réutilise la taxonomie TerraSARX et étend celle de TELEIOS (2016).

L'approche de Keßler & Farmer (2015) utilise à la fois des sources de données externes et des relations temporelles. Elle s'appuie sur le concept de *prisme spatio-temporel* ("space-time prism") qui décrit les caractéristiques de trajectoires. Un outil très connu pour l'analyse spatiale est le *buffer* ; il s'agit d'une zone située autour d'une région d'intérêt, mesurée dans une certaine unité spatiale. Le *prisme spatio-temporel* est similaire à un *buffer* auquel on a ajouté une dimension temporelle. Une fois la zone autour de la caractéristique d'intérêt définie, il est possible d'identifier d'autres propriétés contenues dans le *prisme spatio-temporel*. Ainsi, dans les trajectoires étudiées par Keßler & Farmer (2015), certains points sont caractérisés par des valeurs temporelles de type *xsd:dateTime*. Les valeurs extraites déterminent si un élément est contenu ou non dans un *prisme spatio-temporel* donné.

Compte-tenu du volume d'information à gérer, la recommandation RDF DataCube du W3C (Brizhinev *et al.*, 2017) suggère de lier les images à des tuiles de telle sorte à faire des assertions sur les tuiles. Dans ce cadre, les tuiles sont des zones carrées géo-localisées définies suite à une décomposition de la surface de la terre à partir d'une grille. Chaque image fournie par Sentinel-2 Single Tile (S2ST) a déjà une tuile. Nous nous appuyons également sur la notion de tuile dans notre processus d'intégration.

## 2.2 Données géographiques et LOD

La collecte et l'intégration de données géographiques produites par une diversité de disciplines est au coeur du projet Digital Earth (Gore, 1998). Rendre les données géographiques disponibles puis interopérables à un niveau sémantique est une préoccupation qui conduit à appliquer les principes du Linked Data (Blázquez *et al.*, 2014), et ainsi exposer, partager, et intégrer les données sur le Web via des URI déréférençables (Heath & Bizer, 2011). Les guides du W3C pour publier des données spatiales sous forme de LOD (Linked Open Data), attirent l'attention sur la représentation des relations spatiales et des systèmes de référence (CRS, pour "Coordinate Reference Systems") (Tandy *et al.*, 2017). De nombreuses ontologies et vocabulaires sont recommandés pour représenter des données raster géospatiales volumineuses dans le LOD (Linked Open Data). Le W3C suggère l'ontologie RDF Data Cube (QB), introduite ci-dessus, combinée à d'autres ontologies standards du W3C et de l'OGC dont SSN (Semantic Sensor Network) [4], OWL-Time [5], SKOS [6], PROV-O [7] et la récente extension de DataCube pour les entités spatio-temporelles, QB4ST [8].

L'OGC a introduit la notion de *données géo-liées* ("geolinked data") pour faire référence aux données liées géographiquement. Dans les premiers travaux, la géométrie était stockée dans un ensemble de données géospatiales séparé, et non directement comme valeur d'attributs. Cette option étant trop contraignante lorsqu'il faut comparer la géométrie de chaque entité, les entrepôts actuels mémorisent ensemble, une représentation RDF de la géométrie et une représentation RDF des entités spatiales. Atemezing (2015) a identifié divers types de géométries (point, ligne ou polygone) et divers outils pour construire une représentation RDF

4. `http://purl.oclc.org/NET/ssnx/ssn`
5. `https://www.w3.org/TR/owl-time`
6. `http://www.w3.org/2004/02/skos/core`
7. `https://www.w3.org/TR/prov-o`
8. `https://www.w3.org/TR/qb4st/`

de la géométrie (comme Geometry2RDF [9] ou TripleGeo [10]). Il modélise aussi quatre vocabulaires pour représenter les CRS, les entités topographiques et leurs géométries. Ces ontologies étendent des vocabulaires existant et offrent deux avantages supplémentaires : une utilisation explicite du CRS identifié par des URI pour la géométrie, et la possibilité de décrire des géométries structurées en RDF. Les données sont publiées comme la base de données française GEOFLA.

Une autre considération à intégrer pour représenter des données d'observations de la Terre est la différence de validité temporelle des données. Certaines données, comme la position des stations météorologiques, des villes et de la plupart des lieux administratifs, et même la couverture terrestre, sont valides pour une très longue période, plus longue que celle de l'application, et peuvent être considérées stables ou statiques. Les flux de données, au contraire, fournissent en continu de nouvelles données à intervalles de temps réguliers. Par exemple, les mesures de température sont données toutes les 3 heures par les bulletins Météo France, et des dizaines d'images satellites et leurs méta-données sont disponibles sur le serveur PEPS chaque jour.

## 3 Modèle d'intégration

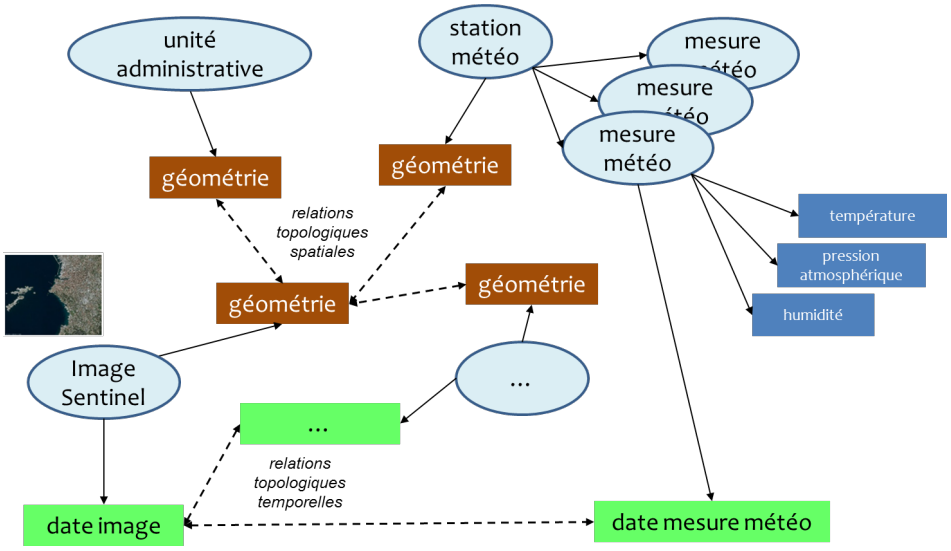



FIGURE *1 – Illustration du rôle pivot de la géométrie et de la datation dans le modèle de données pour associer les données par des relations topologiques.*

Le modèle retenu pour intégrer des données aux métadonnées des images repose sur l'hypothèse que ces données sont géolocalisées et datées. Pour chacune, on doit connaître a minima un point déterminé par sa latitude / longitude, ou une zone géolocalisée (appelée « géometrie » dans la Figure 1). C'est par le biais de la géométrie que des données peuvent être associées à une même image ou à une zone dans une image. Par exemple sur la Figure 1, les mesures météorologiques (humidité, température, pression) ainsi que les unités administratives (villes, régions, etc.) sont géolocalisées et couvrent une superficie plus ou moins grande. Pour comparer les géométries des données à lier, il faut les ramener à une même unité, ce qui permet de savoir quelle zone de l'image est concernée ou décrite par ces données. Les images étant datées, dès lors que les autres données sont aussi datées, comme les mesures météorologiques, il est possible de les lier par des relations temporelles, et de savoir par exemple

9. https://github.com/boricles/geometry2rdf
10. https://github.com/GeoKnow/TripleGeo

quelles sont les températures relevées sur la zone couverte par l'image durant les 3 jours qui ont précédés la prise d'image.

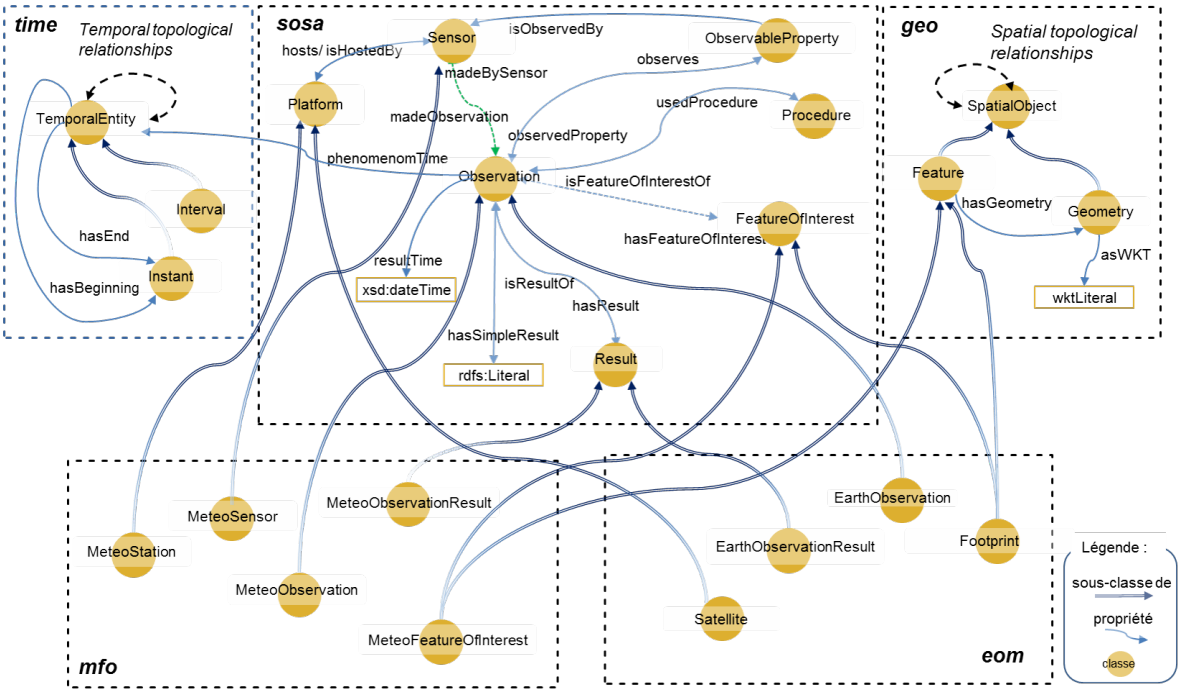



FIGURE 2 – *Le modèle d'intégration de données.*

Nous détaillons à présent le modèle conçu comme support à cette approche. Pour permettre l'intégration de divers types de données, ce modèle, présenté sur la Figure 2 comporte une partie générique (les trois cadres *time*, *sosa* et *geo* de la partie haute de la figure), composée de classes et de propriétés tirés de vocabulaires existants (respectivement OWL-Time, SOSA, GeoSPARQL), et une partie spécifique aux données à intégrer (cadres de la partie basse de la Figure 2). La partie spécifique comprend a minima un vocabulaire pour décrire les métadonnées d'images (cadre *eom* pour "Earth Observation Model") et autant de vocabulaires que de types de données à intégrer (un seul vocabulaire est mentionné ici, *mfo* "pour MétéoFrance Observation", représentant les données météorologiques).

## 3.1 Les vocabulaires réutilisés

Pour faciliter le partage de connaissances, une bonne pratique dans la construction d'ontologies consiste à identifier des ontologies existantes à réutiliser, à les connecter ou les étendre. Nous commençons donc par décrire les vocabulaires sur lesquels s'appuie la partie générique de notre modèle : OWL-Time, SOSA et GeoSPARQL, désignés par la suite respectivement par les préfixes *time*, *sosa* et *geo*. D'autres vocabulaires peuvent être réutilisés en fonction des données à intégrer pour la partie spécifique.

### 3.1.1 Vocabulaire pour représenter méta-données d'images et données de capteurs

Un grand nombre de modèles ont été développés pour représenter des méta-données d'observations par les communautés des sciences de la Terre et de l'environnement. O&M (Observations and Measurements), développé par le groupe SWE de l'OGC, fournit une vision utilisateur des observations. L'OGC Sensor Observation Service and Web Feature Service en a fourni une implémentation XML (spécification XSD). Plus général, DCAT est un vocabulaire recommandé par le W3C pour publier des métadonnées de catalogues sur le Web et disponible au format RDF (Maali & Erickson, 2014).

Pour faciliter l'intégration de données provenant de différents types de capteurs, l'ontologie SSN [11] décrit des réseaux de capteurs numériques ainsi que les fonctionnalités, propriétés et mesures de ces capteurs. Construite comme une sous-couche de l'ontologie formelle DOLCE-UltraLite (DUL), SSN est générique au sens où elle ne définit pas toutes les propriétés spécifiques à des capteurs et des observations particuliers. En revanche, on peut la lier à des ontologies ou des vocabulaires propres à un domaine tels que l'ontologie SWEET (Semantic Web for Earth and Environmental Terminology) qui décrit des données environnementales (Raskin, 2006). Pour faciliter sa réutilisation, et éviter les risques d'incompatibilité notamment avec les déclarations DUL, une version modulaire de SSN a été publiée. Elle est composée de plusieurs ontologies couvrant le même domaine, mais avec une portée sémantique différente pour s'adapter à des contextes d'utilisation différents. SOSA (Sensor, Observation, Sample, and Actuator) est le module ontologique noyau de la nouvelle version de SSN ; à ce titre, il regroupe les termes centraux de l'ontologie. Cox (2017) proposent des alignements entre différents modèles de données d'observations, dont SOSA, à l'aide de l'ontologie PROV-O, pour faciliter l'intégration des données produites avec chacun d'eux. Les axiomes d'alignement sont fournis en RDF dans un module de la spécification SOSA/SSN. Ils s'appuient notamment sur le fait qu'une observation dans le vocabulaire SOSA (*sosa:Observation*) est considérée comme une activité (*prov:Activity*) de capteur, i.e. tout processus (calcul, simulation, interprétation, etc.) appliquant une procédure pour obtenir une estimation d'une valeur de propriété. La classe *sosa:Observation* y est ainsi définie comme une sous-classe de *prov:Activity*.

Nous avions dans un premier temps fait le choix de DCAT et SSN pour représenter respectivement les catalogues de méta-données d'images et les données issues de stations météorologiques (Arenas *et al.*, 2016a,b). SOSA étant pertinent pour une vaste gamme d'applications, dont l'imagerie satellite, nous l'avons finalement retenu pour représenter à la fois les méta-données d'image et les données météorologiques.

### 3.1.2 Vocabulaire pour représenter la composante spatiale des données

Pour représenter l'information géospatiale, une des normes les plus connues est GeoSPARQL, un standard de l'OGC qui définit une petite ontologie pour la représentation de caractéristiques, de relations et de fonctions spatiales (Kolas *et al.*, 2013; Battle & Kolas, 2012; Perry & Herring, 2012). Le cadre *geo* de la Figure 2 présente les principales classes de GeoSPARQL. La classe *geo:Feature* représente toute entité du monde réel ayant une empreinte spatiale. Cette empreinte est décrite par une "géométrie" (point, polygone, etc.), instance de la classe *geo:Geometry*. Une entité est liée à sa géométrie via la propriété *geo:hasGeometry*. Les coordonnées d'une géométrie sont décrites via la propriété *geo:asWKT*. Il est possible de lier GeoSPARQL à une ontologie de n'importe quel domaine en spécialisant la classe *geo:Feature* par une classe de l'ontologie de domaine considérée.

Parmi les alternatives à GeoSPARQL, GeoRDF sert à représenter des propriétés géographiques de points telles que la latitude, la longitude, l'altitude (en utilisant WGS84 comme référentiel) ; GeoOWL permet de représenter l'empreinte d'objets spatiaux plus complexes (lignes, rectangles, polygones) ; stSPARQL est une autre extension de SPARQL conçue pour interroger des graphes RDF spatio-temporels (Koubarakis & Kyzirakos, 2010; Bereta *et al.*, 2013). Nous avons retenu GeoSPARQL car il offre la possibilité de raisonner sur des entités (*geo:Feature*) ayant une forme géométrique (*geo:Geometry*), à savoir proposer des relations entre des entités sur la base des relations entre leurs géométries (inclusion, recouvrement, etc). GeoSPARQL assure l'expression de relations topologiques spatiales binaires sous forme de propriétés des *geometries* et des *features*, et de fonctions servant à les comparer sur la base des propriétés. Parmi les fonctions spatiales de GeoSPARQL, on peut citer *equals*, *disjoint*, *intersects*, *touches*, *within*, *contains*, *overlaps*, et *crosses*. Grâce à des requêtes utilisant ces fonctions, il est possible de comparer, à la volée, deux "*geometries*" ou deux "*features*", ou de comparer une ressource géolocalisée à une position fournie explicitement. La requête GeoSPARQL ci-après recherche les "objets" situés dans (fonction GeoSPARQL *sfWithin*) un polygone (un triangle) défini par les coordonnées de ses trois sommets :

11. `https://www.w3.org/TR/vocab-ssn/`

```
select ?obj_loc
WHERE {
?obj_loc a geo:Feature .
?obj_loc geo:hasGeometry ?geo_obj .
FILTER ( geof:sfWithin( ?geo_obs ,
        "POLYGON ((-3.562096 42.073807,
                -0.44234999999998 42.476536,
                -0.78979299999997 43.973148))"^^geo:wktLiteral) )
}
```

### 3.1.3 Vocabulaire pour représenter la composante temporelle des données

Parmi les modèles de données et langages de requêtes du web sémantique permettant de représenter la composante temporelle de données, on trouve OWL-Time, le standard du W3C, et SWRL Temporal Ontology. Le cadre *time* de la Figure 2 présente les principales classes de OWL Time [12]. La classe *time:TemporalEntity* représente toute entité ayant une temporalité, i.e. un début (propriété *time:hasBeginning*) et une fin (propriété *time:hasEnd*), et donc une durée (propriété *time:hasDuration*). Une entité dont la durée n'est pas nulle ("début" et "fin" sont différents) est un *intervalle*, sinon il s'agit d'un *instant* (*time:Interval* et *time:Instant* étant des spécialisations de *time:TemporalEntity*). Les entités temporelles peuvent être liées par des relations binaires comme *meets*, *overlaps*, *during* issues de l'algèbre des intervalles d'Allen et servant de base pour les raisonnements spatio-temporels.

Le temps a été associé aux triplets RDF à diverses fins. Ainsi, pour donner à un triplet RDF une dimension temporelle représentant sa période de validité (*valid-time*), notion issue des Bases de Données temporelles, qui s'oppose à celle de "date de transaction" (*transaction-time*) indiquant quand un élément a été enregistré dans la base, Koubarakis & Kyzirakos (2010); Bereta *et al.* (2013) proposent le langage stRDF. Un fait y est représenté par un quadruplet <*s,p,o*>*:t* où *t* est une variable de type *xsd:dateTime* (un instant) ou *strdf:period* (un intervalle de temps). stSPARQL, le langage de requête associé à stRDF, supporte des fonctions permettant d'établir des relations temporelles et ainsi de rechercher des faits qui se sont produits entre deux dates. Des représentations plus complexes sont également proposées pour représenter les évolutions d'objets au cours du temps. Ainsi le modèle Continuum s'appuie sur une ontologie des *fluents* et enrichit GeoSPARQL pour modéliser les changements qui s'opèrent sur des entités spatiales (Harbelot *et al.*, 2014).

Pour représenter la date des observations, SOSA intègre une dimension temporelle basée sur OWL-Time. Nous utilisons donc aussi ce vocabulaire.

## 3.2 Les vocabulaires spécifiques du modèle

Nous avons donc adopté SOSA pour décrire les données d'observations (méta-données d'image, observations météorologiques, etc.). La propriété *sosa:phenomenomTime* ayant *sosa:Observation* pour domaine et *time:TemporalEntity* pour co-domaine sert à représenter la période de temps durant laquelle le résultat d'une observation a été obtenu.

Pour représenter les relations spatiales, nous complétons notre modèle avec GeoSPARQL. Nous partons du principe que pour toute observation géolocalisée, son "*point d'intérêt*" (*FeatureOfInterest*) correspond à la zone observée ; cette zone étant caractérisée par une géométrie, elle est également représentée comme un *geo:Feature*. Ainsi, pour les différents types de sources d'observations issues de capteurs, notre approche consiste à définir un nouveau vocabulaire, avec une espace de nom spécifique, approprié à la source des données. Ce vocabulaire comporte au moins une classe qui spécialise à la fois *sosa:FeatureOfInterest* et *geo:Feature*.

Les deux cadres en bas de la Figure 2 correspondent à deux vocabulaires développés selon cette approche pour représenter les méta-données d'images satellites pour l'un (vocabulaire *eom*), les observations de stations météorologiques pour l'autre (vocabulaire *mfo*). Ces vocabulaires sont présentés dans la partie 4.1.1. Dès lors qu'on est capable de représenter la composante spatiale d'une image comme instance de *geo:Feature*, on est capable de la lier à d'autres informations ayant une composante spatiale également définie comme une *geo:Feature*. De même, on peut lier par une relation temporelle un enregistrement de méta-données d'image avec des mesures observées par ailleurs, telles que les données météo, ou avec des périodes d'intérêt (par exemple "une semaine après la prise de l'image").

12. https://www.w3.org/TR/owl-time/ (10/2017)

# 4 Entrepôts RDF pour les données d'observations de la Terre

## 4.1 Les sources de données et leurs composantes spatiale et temporelle

Nous précisons ici comment nous avons adapté les ontologies réutilisées pour concevoir un vocabulaire adapté à chaque source de données. Parmi les données géolocalisées, nous avons distingué les *données dynamiques*, dont la validité est fournie par la composante temporelle, et les *données statiques*, qui ne sont pas datées a priori ou auxquelles sont associés de très longs intervalles de temps. Nous présentons à présent les deux (sous-)modèles résultant et leur instanciation pour enrichir les méta-données d'images. Il est à noter que la propriété *time:hasTime* de OWL-Time est définie comme "un prédicat générique pour associer une entité temporelle à n'importe quoi". Donc même si notre modèle ne le mentionne pas, toute entité décrite par une URI peut être estampillée par une composante temporelle (une date ou un intervalle de temps).

### 4.1.1 Les observations (données dynamiques)

Le modèle est spécialisé dans 2 modules dédiés à chacune des sources d'observations issues de capteurs décrites ici, à savoir les images satellites et les observations météo.

#### *4.1.1.1 Les méta-données d'images*

Dans le projet SparkInData, nous utilisons des enregistrements de méta-données d'images Sentinel [13]. La périodicité de Sentinel-1 est de douze jours, celle de Sentinel-2 étant de cinq jours. Les enregistrements de méta-données sont obtenus au format GeoJSON à partir de RESTO, un service géré par le CNES (Gasperi, 2014). L'API RESTO permet de spécifier les paramètres à retrouver, à savoir des métadonnées d'enregistrements comme la couverture nuageuse, l'intervalle de temps, la zone géographique d'intérêt, etc. Nous collectons ces informations toutes les nuits. L'URL suivante fait appel à RESTO pour retourner tous les enregistrements de métadonnées de la collection S2ST pour la France, captés entre le 19/09/2017 23:00 et le 25/09/2017 00:00 : `https://peps.cnes.fr/resto/api/collections/S2ST/search.json?q=France\&startDate=17-09-19T23:00:00\&completionDate=2017-09-25T00:00:00`.

Les méta-données d'images sont obtenues sous forme de fichiers GeoJSON ; elles sont converties en RDF et représentées à l'aide du vocabulaire *eom* comme des "observations de la Terre", i.e. des instances de *eom:EarthObservation* (cadre *eom* de la Figure 2). Cette classe étant une sous-classe de *sosa:Observation*, la composante spatiale des observations est fournie via la classe *eom:Footprint*, spécialisation de *geo:Feature* et *sosa:FeatureOfInterest* : un *footprint* est décrit par un polygone fermé (une géométrie) délimitant la zone terrestre couverte par l'image. La classe *eom:Footprint* étant une sous-classe de *geo:Feature*, connaissant l'empreinte des images (leur *footprint*), il est facile d'identifier les caractéristiques d'un autre type de données qui recouvrent la position des images. La *dimension temporelle* d'un enregistrement de méta-données d'image indique le moment où l'image a été prise. Elle est fournie via la propriété *sosa:phenomenomTime* de l'observation correspondante.

Le module est complété par ailleurs pour fournir des données propres aux images satellites telles que le domaine spectral de l'image (infrarouge, UV, etc.), la direction orbitale du satellite (ascendante, descendante) ou encore la couverture nuageuse de l'image.

#### *4.1.1.2 Les observations météorologiques*

Comme données contextuelles dynamiques, nous utilisons les informations météo de *SYNOP Meteo France* [14]. Ces observations sont réalisées toutes les trois heures par chacune des 62 stations françaises. Un fichier séparé contient la liste des stations avec leur position respective (un point fixe repéré par ses coordonnées géographiques). Selon une approche très similaire à celle de Lefort *et al.* (2012), nous utilisons des classes tirées de SOSA que nous spécialisons pour définir des classes propres au domaine de la météorologie. Dans un premier temps, nous sommes restés au plus près du schéma de la source MétéoFrance utilisée. Nous envisageons de faire évoluer ce modèle en intégrant le vocabulaire SWEET (Raskin, 2006), devenu un standard pour les données météo.

Nous représentons une station météo comme une instance de la classe *mfo:MeteoStation*, une sous-classe de *sosa:Platform*. Les capteurs fonctionnant sur une station météo sont représentés comme

13. `https://sentinel.esa.int/web/sentinel/missions/` (07/2016)

14. `https://donneespubliques.meteofrance.fr/` (07/2016)

instances de *mfo:MeteoSensor*, sous-classe de *sosa:Sensor*. La position géographique où s'effectuent les mesures est représentée comme instance de la classe *mfo:MeteoFeatureOfInterest*, une sous-classe de *sosa:FeatureOfInterest*.

Comme pour *eom*, le module *mfo* a été complété pour représenter les variables fournies avec les relevés météo ; il y a 28 variables par relevé, mais beaucoup d'entre elles ne sont pas renseignées. Nous avons ainsi instancié la classe *sosa:ObservableProperty* avec notamment des individus représentant la température (*mfo:Temperature*), l'humidité (*mfo:HUMIDITY*) ou la vitesse du vent (*mfo:WIND_SPEED*), variables dont une valeur est relevée toutes les trois heures. Le module est donc enrichi régulièrement par de nouvelles instances (tous les 5 ou 12 jours pour les images satellites, quotidiennement pour les relevés météo). Le code ci-après est un extrait de la représentation RDF d'un relevé de température réalisé le 04/02/2018, entre 3h et 15h :

```
g-mfo:obs_07005_20180204150000_tminsol a mfo:Observation .
g-mfo:obs_07005_20180204150000_tminsol sosa:phenomenonTime
                                       g-mfo:interval_1517713200-1517756400 .
g-mfo:interval_1517713200-1517756400 a time:Interval .
g-mfo:interval_1517713200-1517756400 time:hasBeginning g-mfo:instant_1517713200 .
g-mfo:interval_1517713200-1517756400 time:hasEnd g-mfo:instant_1517756400 .
g-mfo:instant_1517713200 a time:Instant .
g-mfo:instant_1517713200 time:inXSDDateTime "2018-02-04T03:00:00"^^xsd:dateTime .
g-mfo:instant_1517713200 time:inXSDDateTimeStamp "1517713200"^^xsd:dateTimeStamp .
g-mfo:obs_07005_20180204150000_tminsol sosa:hasResult
                                       g-mfo:obs_07005_20180204150000_tminsol_result .
g-mfo:obs_07005_20180204150000_tminsol_result a mfo:Result .
g-mfo:obs_07005_20180204150000_tminsol_result a qudt-1-1:QuantitativeValue .
g-mfo:obs_07005_20180204150000_tminsol_result qudt-1-1:unit qudt-unit-1-1:Kelvin .
g-mfo:obs_07005_20180204150000_tminsol_result qudt-1-1:numericValue "274.75"^^xsd:double .
g-mfo:obs_07005_20180204150000_tminsol sosa:hasFeatureOfInterest
                                       <http://melodi.irit.fr/lod/mfo/foi_07005> .
g-mfo:obs_07005_20180204150000_tminsol sosa:observedProperty g-mfo:Temperature .
g-mfo:obs_07005_20180204150000_tminsol sosa:madeBySensor g-mfo:station_07005_Thermometer .
g-mfo:obs_07005_20180204150000_tminsol sosa:usedProcedure g-mfo:procedure_tminsol .
```

### 4.1.2 Les données supports (données statiques)

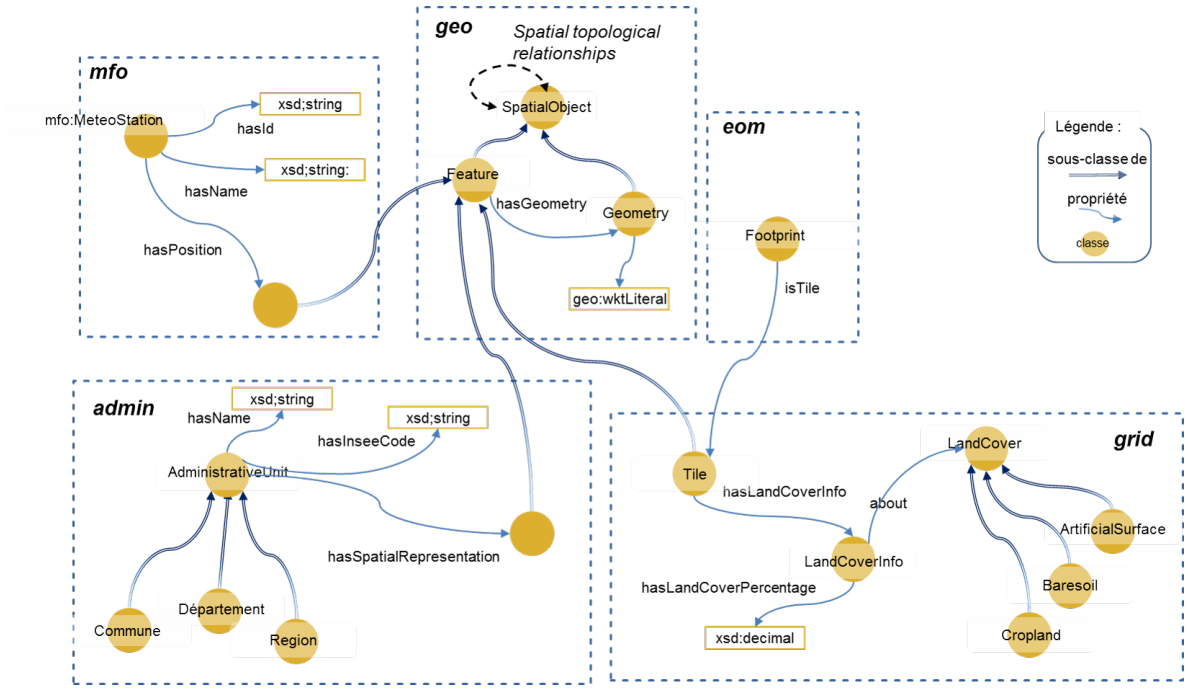



FIGURE *3 – Les vocabulaires* admin *et*grid *intégrant des données "statiques".*

Le modèle peut être enrichi pour représenter des données dites statiques, i.e. qui ont une composante spatiale mais dont la représentation de la composante temporelle n'est pas nécessaire car leur valeur ne varie pas dans l'échelle de temps considérée. Nous avons créé deux modules de ce type, *admin* et *grid*, présentés sur la Figure 3. Pour mettre en relation ces données avec les données dynamiques

via des relations spatiales, un concept de chacun de ces modules spécialise la classe *geo:Feature*. Les classes de ces modules sont instanciées une fois pour toutes.

#### *4.1.2.1 Les tuiles d'images*

Les images Sentinel ont des caractéristiques différentes en fonction du capteur qui les a prises. En septembre 2016, l'ESA a commencé à diffuser des images Sentinel 2 sous forme de paquets de tuiles simples (images S2ST), chaque tuile représentant une zone fixe de la surface du globe de taille approximative de 100 x 100 km. Une image S2ST correspond donc à un fragment d'image S2. L'intérêt par rapport à une image S2 est que l'utilisateur peut mieux sélectionner la surface qui l'intéresse et ne télécharger que l'information souhaitée. Un fichier S2ST est aussi moins volumineux : il peut faire environ 500 Mo alors que celui d'une image S2 avant tuilage peut faire plus de 3 Go.

La grille de l'ESA décrivant le tuilage de la surface terrestre (Cf. Figure 4), i.e l'empreinte de chaque tuile, est fournie dans un fichier Grid au format KML [15]. Nous avons converti ce fichier en RDF à l'aide du vocabulaire *grid* décrit en bas de la Figure 3 : chaque tuile est représentée comme une instance de la classe *Tile*, spécialisation de *geo:Feature*, dont la propriété *geo:hasGeometry* correspond à son empreinte. Le module *eom* (Cf Section 4.1.1) a été enrichi avec la propriété *eom:isTile* pour associer une tuile au *footprint* d'une image. Ainsi en reliant des données à une tuile, on associe indirectement ces données à toutes les images associées à cette tuile.

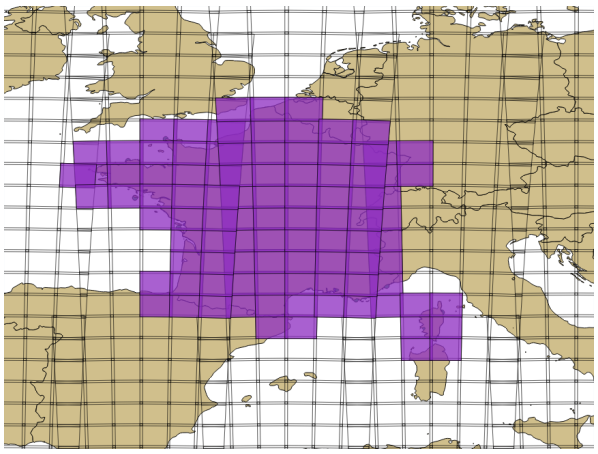


FIGURE *4 – Les tuiles de l'ESA définies pour la France métropolitaine*

#### *4.1.2.2 La couverture terrestre*

Pour associer des données aux tuiles, nous avons utilisé une autre source de données statiques, le GLC-SHARE (Global Land Cover SHARE) produit par le FAO, qui donne des informations sur la couverture terrestre. Cette dernière est définie à partir d'une nomenclature permettant de classer les zones en fonction du type d'occupation des sols ou du type de surface (surface artificielle, terre cultivée, zone forestière, etc.). Les données du GLC-SHARE sont fournies sous forme d'une image au format TIFF dont chaque pixel correspond à une échelle spatiale approximativement de 1 $km^2$. La valeur d'un pixel est un entier indiquant la classe la plus fréquente pour la zone couverte par le pixel. Nous avons donc calculé la composition de la couverture terrestre de chaque tuile de l'ESA de la France. Avec le vocabulaire *grid*, une tuile est liée à un ensemble de graphes RDF (via la propriété *hasLandCoverPercentage*) décrivant chacun le pourcentage d'une classe GLC-SHARE (*CropLand*, *Baresoil*, etc.) sur la surface couverte par la tuile.

15. `https://sentinel.esa.int/web/sentinel/missions/sentinel-2/news/-/article/sentinel-2-tiling-grid-updated`

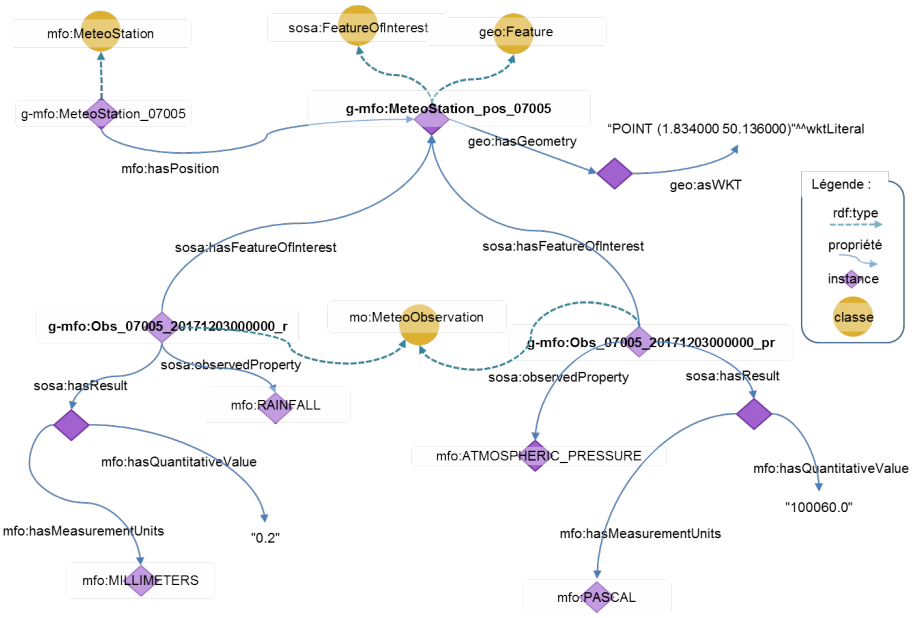



FIGURE 5 – *Instanciation du sous-modèle "mfo" : le "feature of interest" des observations d'une station météo est le lieu où se trouve la station.*

#### *4.1.2.3 Les stations météorologiques*

Les informations sur les stations météo, tirées du fichier fourni par *SYNOP Meteo France*, fournissent entre autre la position géographique des mesures météo. Le modèle *mfo* (Cf Section 4.1.1) a ainsi été complété notamment avec la propriété *hasPosition* dont le codomaine est une spécialisation de *geo:Feature*. Le code suivant est un extrait de la représentation RDF de la station météorologique d'Embrun :

```
g-mfo:MS_07591    rdf:type    mfo:MeteoStation .
g-mfo:MS_07591    sosa:hosts  g-mfo:Sensor_07591_humidity .
g-mfo:MS_07591    sosa:hosts  g-mfo:Sensor_07591_temperature .
g-mfo:MS_07591    mfo:hasId   "07591"^^xsd:String .
g-mfo:MS_07591    mfo:hasName "EMBRUN"^^xsd:String .
g-mfo:MS_07591    mfo:hasPosition g-mfo:MS_pos_07591 .
g-mfo:MS_pos_07591   rdf:type  geo:Feature .
g-mfo:MS_pos_07591   geo:hasGeometry g-mfo:MS_geo_07591 .
g-mfo:MS_geo_07591 rdf:type    geo:Geometry .
g-mfo:MS_geo_07591 geo:asWKT "POINT ((6.502333 44.565667))^^geo:wktLiteral" .
```

La Figure 5 est un graphe d'instances illustrant (partiellement) la représentation d'observations météo avec le vocabulaire *mfo* : 2 observations mesurant respectivement les précipitations et la pression atmosphérique, ont été réalisées par la même la station météo. La figure met en évidence qu'une seule URI (*g-mfo:MeteoStation_pos_07005* sur l'exemple) est utilisée pour représenter à la fois la position de la station (propriété *mfo:hasPosition*) et le lieu de chacune des observations (propriété *sosa:hasFeatureOfInterest*).

#### *4.1.2.4 Les unités administratives*

Afin de lier les observations de la Terre à des entités administratives (villes, départements, régions, etc.) françaises à partir de leur position géographique (point ou polygone), nous avons enrichi le modèle avec le module *admin* (Cf. Figure 3). La classe *AdministrativeUnit* est caractérisée par une propriété *hasSpatialRepresentation* dont le co-domaine est une sous-classe de *geo:Feature*.

Le modèle a été instancié avec des données provenant de la plate-forme ouverte des données publiques française « data.gouv.fr »[16]. Ces données étaient initialement fournies au

16. https://www.data.gouv.fr

format shapefile. Les entités administratives de notre triplestore (instances de *AdministrativeUnit*) sont alignées via la propriété *owl:sameAs* avec les territoires français de l'ontologie géographique de INSEE [17] ; cette ontologie ne contenant pas d'information géométrique, il n'était pas possible de l'utiliser pour établir des relations spatiales. Ci-après un exemple de graphe RDF, celui de la région 54 :

```
admin:region_54 a admin:AdministrativeUnit .
admin:region_54 a admin:Region .
admin:region_54 owl:sameAs <http://id.insee.fr/geo/region/54> .
admin:region_54 admin:hasInseeCode "54"^^xsd:String .
admin:region_54 admin:hasName "Poitou-Charentes"^^xsd:String .
admin:region_54 a geo:Feature .
admin:region_54 geo:hasGeometry l_admin:region_54_geo .
admin:region_54_geo geo:asWKT "MULTIPOLYGON(((
-1.0988062299633785 45.64032288975508, ...
-1.0988062299633785 45.64032288975508))...)"^^wkt:Literal .
```

## 5 Intégration de données RDF via leurs relations spatiales

Chacun des modules des Figures 2 et 3 est décrit dans un fichier OWL spécifique. Le modèle global est ainsi composé de 4 ontologies [18]. Le calcul des relations spatiales et temporelles permet de relier les données de chacun de ces entrepôts. En liant des éléments spatiaux aux tuiles de l'ESA, les connaissances s'y rapportant sont extrapolées au niveau des images. Par exemple, sachant que les images $[img1, img2, img3]$ partagent la tuile $tile_1$, et que cette tuile recouvre $adminUnit_i$, il est possible d'inférer que $[img1, img2, img3]$ recouvrent aussi $adminUnit_i$.

Pour calculer une relation spatiale R (appartenance ou non-intersection, par exemple) entre deux ensembles de données *S* et *T* d'une base de connaissances, nous comparons chaque entité *s* de *S* à chaque entité *t* de *T*, pour vérifier l'existence de *R(s,t)*. Ceci est réalisé à l'aide d'une requête comme celle ci-dessous qui recherche l'ensemble des triplets < s, geo:sfIntersect, t > tels que les *géométries* (les empreintes géographiques) de *s* et *t* ont une intersection non vide :

```
CONSTRUCT {?s geo:sfIntersects ?t .}
WHERE {
 GRAPH <S> {?s geo:hasGeometry/geo:asWKT ?s_geometry.}
 GRAPH <T> {?t geo:hasGeometry/geo:asWKT ?t_geometry.}
 FILTER( geof:sfIntersects(?s_geometry , ?t_geometry) )
 }
```

Cette approche, consistant à faire le produit cartésien de deux ensembles, est de complexité quadratique si bien que, dès que les jeux de données à combiner sont volumineux, calculer ces relations au moment de l'interrogation, peut être extrêmement coûteux en temps et inacceptable pour les applications temps-réel. C'est pourquoi il est préférable de les pré-calculer. Nous avons distingué le cas des données à composante spatiale seule (statiques) de celles ayant en plus une composante temporelle (dynamiques).

### 5.1 Intégration des données statiques

Pour les données statiques, il est raisonnable d'envisager d'enregistrer dans l'entrepôt RDF les relations spatiales entre les données RDF correspondantes, si ces jeux de données ont une taille raisonnable. Si le volume est trop conséquent, il est nécessaire d'avoir recours à des techniques d'optimisation. Ainsi, en nous appuyant sur l'indexation spatiale fournie par le tuilage des images S2ST, nous avons calculé les relations entre les données de chacun des jeux

17. http://rdf.insee.fr/def/index.html

18. http://melodi.irit.fr/ontologies/eom.owl http://melodi.irit.fr/ontologies/mfo.owl http://melodi.irit.fr/ontologies/grid.owl http://melodi.irit.fr/ontologies/administrativeUnits.owl

que nous venons de présenter : stations météo de *mfo*, unités administratives de *admin* et tuiles de *grid*. Ainsi, il est possible de lier les stations météorologiques aux images en calculant uniquement les relations spatiales entre les stations météo et les tuiles de l'ESA. La requête SPARQL ci-dessous interroge les jeux de données instanciant les modèles *eom*, *grid* et *admin*. Elle retrouve les images S2ST avec une couverture nuageuse inférieure à 10%, collectées à un moment donné, et dont l'empreinte géographique couvre (propriété *geo:sfContains*) une zone géographique particulière et ayant un type d'occupation du sol particulier. L'utilisateur a défini une date qui, comme la couverture nuageuse et le type d'occupation du sol, est utilisée pour filtrer les données pertinentes (Cf. les deux dernières clauses FILTER).

```
select distinct ?s2st_result ?tileId
{
?s2st a md:EarthObservation .
?s2st md:featureOfInterest ?fi .
?s2st md:result ?s2st_result .
?s2st_result md:product ?s2st_product.
?s2st_product md:cloudCoverPercentage ?s2st_cloudCover .
?s2st md:phenomenomTime ?s2st_period .
?s2st_period time:hasBeginning ?s2st_time .
?fi md:isTile ?tile .
?tile geo:hasGeometry ?tile_geo .
?tile_geo geo:sfContains ?admin_geo .
?adminUnit admin:hasSpatialRepresentation ?admin_sr .
?admin_sr geo:hasGeometry ?admin_geo .
?adminUnit admin:hasName ?adminUnitName .
?adminUnitName bif:contains 'Alpes' .
?tile grd:hasLandCoverInfo ?tile_lc .
?tile_lc grd:hasLandCoverPercentage ?lc_perc .
?lc_perc grd:about ?lc_class .
?lc_perc grd:percentage_value ?lc_value .
?tile grd:tileId ?tileId .
FILTER (?s2st_time = ?user_time)
FILTER (?lc_value>1 and ?lc_class=grd:ArtificialSurface
            and ?s2st_cloudCover<10)
}
```

Nous avons testé deux méthodes pour calculer les relations spatiales et créer les triplets RDF les représentant. Nous venons d'exposer la première qui met en oeuvre des requêtes SPARQL incluant des fonctions GeoSPARQL. La seconde a consisté à développer un programme en Python, à l'aide du module Shapely qui permet de faire des opérations spatiales sur des données au format WKT. Avec la solution Python, il faut 0,50s (sans optimisation particulière) pour calculer toutes les relations spatiales entre 1 élément d'un jeu de données *S* et 50 éléments d'un jeu de données *T*, alors qu'il faut 11s avec GeoSPARQL pour calculer une seule relation entre 1 élément de *S* et 50 éléments de *T*.

### 5.2 Intégration des données ayant une composante temporelle

Il est possible d'établir des relations temporelles entre un enregistrement de méta-données d'image et des données ayant une indication temporelle comme les relevés météo, ou des périodes d'intérêt définies par l'utilisateur. A défaut de disposer d'entités qui servent de référentiel temporel (le pendant des tuiles de l'ESA pour la composante spatiale), nous exploitons l'intervalle de temps défini par un utilisateur lors de la recherche d'images, comme un buffer temporel fournissant un contexte aux enregistrements de méta-données.

La Figure 6 illustre ce principe. Une interface permet à l'utilisateur de dessiner un rectangle pour spécifier des contraintes spatiales et de saisir une période de temps pour définir un empan temporel. A partir de ces informations, le système génère une requête GeoSPARQL qui permet de retrouver les observations météo, les métadonnées et les images dont la composante temporelle recouvre l'empan temporel et dont la localisation vérifie les contraintes spatiales (plus précisément, l'empreinte de ces images a une intersection non nulle avec le rectangle choisi par l'utilisateur). Les images recherchées sont des images Sentinel 2 "classiques", qui ne sont pas liées à une unique tuile comme les S2ST, et donc pour lesquelles

les relations spatiales n'ont pas été pré-calculées. La requête SPARQL ci-dessous sélectionne des enregistrements de méta-données d'images dont les mesures météorologiques associées respectent deux contraintes :

1. elles proviennent de stations situées dans l'empreinte de l'image : `?img_geo` étant la géométrie d'une image et `?ws_geo` celle d'une station météo, la géométrie de l'image doit contenir celle de la station météo, ce qu'indique la fonction `geof:sfContains` dans le filtre `FILTER(geof:sfContains(?img_geo,?ws_geo))`;
2. elles ont été collectées durant une période donnée; l'utilisateur a défini une période d'intérêt d'une semaine après la prise d'image (24x7=168 heures, 'PT168H'), qui est utilisée au sein des deux dernières clauses `FILTER`.

```
SELECT  ?img ?obs_time ?humidityDataVal ?humidityUnits
WHERE{
?img a eom:EarthObservation .
?img eom:featureOfInterest ?img_foi .
?img_foi geo:hasGeometry ?img_geo .
?img eom:phenomenonTime ?img_temp .
?img_temp time:hasEnd/time:inXSDDateTime ?img_end .
?ws a mfo:MeteoStation .
?ws mfo:hasPosition ?ws_pos .
?ws_pos geo:hasGeometry ?ws_geo .
?ws sosa:hosts ?sensor .
?sensor sosa:observes mfo:HUMIDITY .
?obs sosa:madeBySensor ?sensor .
?obs sosa:hasResult ?humidity .
?humidity mfo:hasQuantitativeValue ?humidityDataVal .
?humidityVal mfo:isClassifiedBy ?humidityUnits .
?obs sosa:phenomenonTime ?obs_time .
?obs_time time:hasEnd/time:inXSDDateTime ?obs_end .
BIND (?img_temp-?obs_time as ?diffDateTime)
FILTER (geof:sfContains(?img_geo,?ws_geo))
FILTER (?diffDateTime<'PT168H'^^xsd:dayTimeDuration)
FILTER (?diffDateTime>'PT0H'^^xsd:dayTimeDuration)
}
```

Faute de place, nous n'avons pas indiqué les éléments de la requête filtrant les images présentes dans la zone spécifiée par l'utilisateur.

La Figure 6 montre l'information météorologique associée à une image faisant partie de la liste retournée par une requête de ce type. L'empreinte de l'image est représentée par un polygone orange. Chaque cercle bleu correspond à une station météorologique. En bas de la carte, figurent des séries temporelles qui représentent l'évolution de la variable *Temperature* pendant la période choisie.

## 6 Conclusion

Nous avons proposé une ontologie pour l'intégration de données d'observation de la Terre et donnés contextuelles, basée sur des relations topologiques spatiales et temporelles. Cette ontologie spécialise des standards, notamment SOSA, GeoSPARQL et OWL-Time. Nous avons défini également un processus d'intégration qui sélectionne, transforme et intègre de données géospatiales hétérogènes (méta-données d'image satellite, données météorologiques, unités administratives, couverture terrestre, etc.). Ce processus s'appuie sur le tuilage des images pour traiter les données ayant une composante spatiale fixe, les relations temporelles, quant à elle, sont calculées à la volée à partir d'une topologie temporelle. Nous avons présenté un cas d'étude basé exploitant des méta-données d'image satellite Sentinel.

Dans la continuité de ces travaux, nous envisageons de considérer des sources propres à un domaine métier pour traiter un cas d'usage (l'agriculture et des rapports bulletins agricoles) et fournir des règles et des fonctionnalités de raisonnement pour faciliter les analyses. Il serait intéressant aussi d'identifier des patrons qui augmenteraient la pertinence de l'image, en s'appuyant sur des règles et du raisonnement. On pourrait par exemple analyser les séries

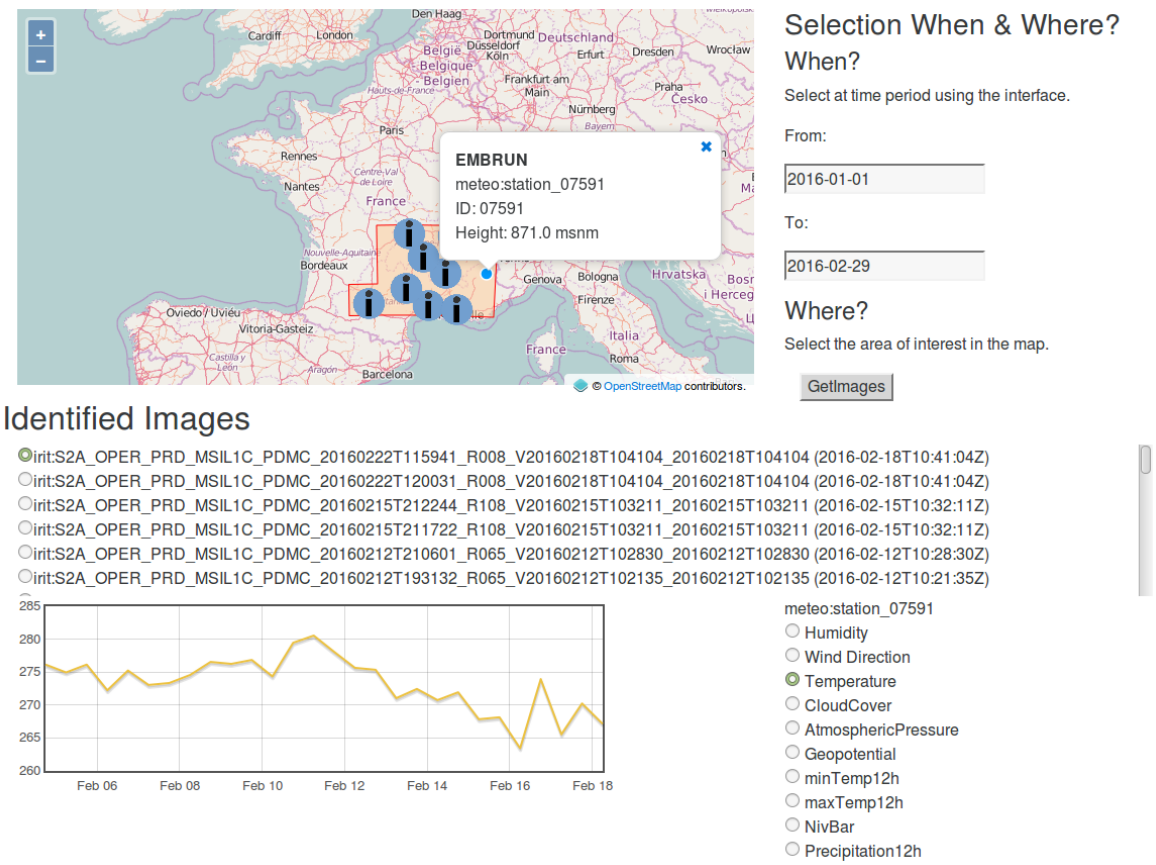



FIGURE 6 – *Affichage des informations météorologiques associées à une image satellite en réponse à la requête.*

temporelles de température et détecter une canicule. Nous envisageons également d'intégrer des alignements automatiques pour mettre en correspondance les métadonnées d'images et les vocabulaires, mais également les données annotées et d'autres sources du LOD.

## Références

ARENAS H., AUSSENAC-GILLES N., COMPAROT C. & TROJAHN C. (2016a). Semantic integration of geospatial data from earth observations. In *Knowledge Engineering and Knowledge Management - EKAW 2016 Satellite Events*, p. 97–100.

ARENAS H., AUSSENAC-GILLES N., COMPAROT C. & TROJAHN C. (2016b). Un modèle pour l'intégration sémantique de données géolocalisées liées à l'observation de la terre. In *Spatial Analysis and Geomatics - Atelier EXtraction de Connaissances à partir de donnéEs Spatialisées (SAGEO-EXCES), Nice*.

ATEMEZING G. A. (2015). *Publishing and consuming geo-spatial and government data on the semantic web*. PhD thesis, Thesis.

BATTLE R. & KOLAS D. (2012). Enabling the Geospatial Semantic Web with Parliament and GeoSPARQL. *Semantic Web*, **3**(October 2012), 355–370.

BERETA K., SMEROS P. & KOUBARAKIS M. (2013). Representation and querying of valid time of triples in linked geospatial data. In *The Semantic Web: Semantics and Big Data*, p. 259–274, Berlin, Heidelberg: Springer Berlin Heidelberg.

BLÁZQUEZ L. M. V., VILLAZÓN-TERRAZAS B., CORCHO Ó. & GÓMEZ-PÉREZ A. (2014). Integrating geographical information in the linked digital earth. *Int. Journal of Digital Earth*, **7**(7), 554–575.

BRIZHINEV D., TOYER S., TAYLOR K. & ZHANG Z. (2017). *Publishing and Using Earth Observation Data with the RDF Data Cube and the Discrete Global Grid System*. Rapport interne, W3C and OGC.

CONSOLE M. & LENZERINI M. (2014). Reducing global consistency to local consistency in ontology-based data access - extended abstract. In *Informal Proceedings of the 27th International Workshop on Description Logics, Vienna, Austria, July 17-20, 2014.*, p. 496–499.

COX S. (2017). Prov ontology supports alignment of observational data (models). In *Modeling and Simulation Society of Australia and New Zealand*, p. 403–409.

ESPINOZA-MOLINA D. & DATCU M. (2013). Earth-observation image retrieval based on content, semantics, and metadata. *IEEE Trans. on Geoscience and Remote Sensing*, **51**(11), 5145–5159.

ESPINOZA-MOLINA D., NIKOLAOU C., DUMITRU C. O., BERETA K., KOUBARAKIS M., SCHWARZ G. & DATCU M. (2015). 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.

GASPERI J. (2014). Semantic Search Within Earth Observation Products Database Based on Automatic Tagging of Image Content. In *Proc. of the Conf. on Big Data from Space*, p. 4–6.

GORE A. (1998). The digital earth. *Australian Surveyor*, **43**(2), 89–91.

HARBELOT B., ARENAS H. & CRUZ C. (2014). un modèle sémantique spatio-temporel pour capturer la dynamique des environnements. In *14 ème conférence Extraction et Gestion des Connaissances*, p. 39 à 54, Rennes, France.

HEATH T. & BIZER C. (2011). *Linked Data: Evolving the Web into a Global Data Space ; Lectures on the Semantic Web: Theory and Technology*. Morgan & Claypool.

KESSLER C. & FARMER C. J. Q. (2015). Querying and integrating spatial-temporal information on the Web of Data via time geography. *Journal of Web Semantics*, **35**, 25–34.

KOLAS D., PERRY M. & HERRING J. (2013). *Getting started with GeoSPARQL*. Rapport interne, OGC.

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*, p. 425–439, Berlin, Heidelberg: Springer Berlin Heidelberg.

LEFORT L., BOBRUK J., HALLER A., TAYLOR K. & WOOLF A. (2012). A linked sensor data cube for a 100 year homogenised daily temperature dataset. In *Proc. of the 5th International Conference on Semantic Sensor Networks - Volume 904*, SSN'12, p. 1–16: CEUR-WS.org.

LEFRANÇOIS M., ZIMMERMANN A. & BAKERALLY N. (2017). A SPARQL extension for generating RDF from heterogeneous formats. In *Proc. Extended Semantic Web Conference (ESWC'17)*, Portoroz, Slovenia.

LENZERINI M. (2011). Ontology-based data management. In *Proc. of the 20th ACM Int. Conference on Information and Knowledge Management*, CIKM '11, p. 5–6, New York, USA: ACM.

MAALI F. & ERICKSON J. (2014). Data Catalog Vocabulary (DCAT).

PERRY M. & HERRING J. (2012). *OGC GeoSPARQL-A geographic query language for RDF data*. Rapport interne, Open Geospatial Consortium.

RASKIN R. (2006). *Guide to SWEET Ontologies*. Rapport interne 9, NASA/Jet Propulsion Lab, Pasadena, CA, USA.

REITSMA F. & ALBRECHT J. (2005). Modeling with the semantic web in the geosciences. *IEEE Intelligent Systems*, **20**(2), 86–88.

SUKHOBOK D., SÁNCHEZ H., ESTRADA J. & ROMAN D. (2017). Linked data for common agriculture policy: Enabling semantic querying over sentinel-2 and lidar data. In *Proceedings of the ISWC 2017 Posters & Demonstrations and Industry Tracks co-located with 16th Int. Semantic Web Conference, Vienna, Austria, Oct. 23-25, 2017*.

TANDY J., VAN DEN BRINK L. & BARNAGHI P. (2017). *Spatial Data on the Web Best Practices, W3C Working Group Note*. Rapport interne, W3C and OGC.

TELEIOS (2016). Virtual Observatory Infrastructure for Earth Observation Data. `http://www.earthobservatory.eu/search/node/ontologies`. Accessed: 2016-03-01.

THE TELEIOS TEAM (2012). Building Remote Sensing Applications Using Scientific Database And Semantic Web Technologies. In *Image Information Mining Conference: Knowledge Discovery from Earth Observation Data*, p. 2–5.

# Besoins ontologiques d'un système d'irrigation intelligent : comparaison entre SSN et SAREF

María Poveda-Villalón[1], Quang-Duy Nguyen[2], Catherine Roussey[2], Christophe de Vaulx[3], Jean-Pierre Chanet[2]

[1] Ontology Engineering Group, Universidad Politécnica de Madrid, Spain mpoveda@fi.upm.es

[2] UR TSCF Irstea, Aubière, France prenom.nom@irstea.fr

[3] Limos, UMR 6158, Aubière, France christophe.de_vaulx@uca.fr

**Résumé** :
L'agriculture de précision est un des leviers importants d'amélioration de l'agriculture en Europe. Cette approche peut notamment s'appliquer à l'irrigation en permettant aux agriculteurs d'économiser l'eau ou de l'utiliser au bon moment sans excès. Cet article présente brièvement les prémices d'un système d'irrigation automatisé pour l'AgroTechnoPôle d'Irstea. Ce système est un système contextuel adaptatif où les ontologies sont utilisées pour résoudre les problèmes d'intégration des données hétérogènes. Ensuite un raisonnement est appliqué sur ces données pour déduire les états des entités et commander le système d'irrigation. L'objectif principal de cet article est d'analyser les ontologies SSN et SAREF produites par deux organismes de normalisation. Nous avons déterminé les besoins de notre système d'irrigation et ensuite nous avons identifié si ces besoins étaient abordés par ces deux ontologies.

**Mots-clés** : Ontologies, Agriculture de précision, Système contextuel adaptatif

## 1 Introduction

Au XXI siècle, les défis économiques et environnementaux obligent les pouvoirs publics et les citoyens à améliorer les processus et les habitudes de vie afin d'économiser les ressources précieuses et rares. L'eau est menacée par le changement climatique et la surexploitation dans de nombreuses régions. Particulièrement dans l'agriculture, domaine qui consomme le plus d'eau au niveau mondial (UNESCO, 2014), il est nécessaire de réduire la consommation d'eau en maintenant les niveaux de production et la qualité des récoltes.

En agriculture, une des activités des agriculteurs est l'observation des phènomènes naturels (météo, croissance des cultures, bioagresseurs...) pour adapter au mieux les pratiques agricoles. Par exemple, l'agriculteur fait le tour des parcelles pour examiner les stades de développement des cultures et l'état des sols afin de décider des actions d'irrigation. Cependant, cette activité d'observation est incertaine en raison des erreurs liées à l'opérateur, ainsi que des éventuels changements rapides de la météo. Ces imprécisions peuvent avoir des impacts significatifs sur les rendements et la qualité des récoltes.

L'agriculture de précision est un facteur essentiel pour l'avenir d'une agriculture écoresponsable en Europe, avec pour principal objectif de surmonter les problèmes de pratiques inadaptées (Schrijver *et al.*, 2016). En agriculture de précision, les technologies numériques sont utilisées pour surveiller le contexte des cultures et ainsi optimiser les pratiques agricoles avec des méthodes adaptées. L'irrigation de précision est une des pratiques clés de l'agriculture moderne. Au cours de la dernière décennie, des sites de démonstration et d'expérimentation pour l'agriculture de précision ont été mis en œuvre pour valider ces approches. Par exemple, sur l'AgroTechnoPôle, situé à Montoldre en France, un système d'irrigation intelligent mettant en œuvre tout le potentiel de l'Internet des Objets est en cours de développement. L'objectif de ce système est de collecter les données d'observation ainsi que le contexte des cultures et de l'exploitation afin d'activer le système d'irrigation au bon moment. Ce type de système connu sous le nom de systèmes contextuels adaptatifs, bénéficie des technologies du Web sémantique comme les ontologies pour harmoniser et intégrer

les données et ensuite raisonner sur ces données. Dans le cas particulier du site de Montoldre, une ontologie est mobilisée pour annoter sémantiquement les données et permettre un raisonnement à base de règles pour déduire l'état du sol et les besoins d'irrigation.

L'objectif de cet article est de : (1) extraire les besoins ontologiques pour un système contextuel d'irrigation et (2) analyser dans quelle mesure ces besoins sont couverts par les deux ontologies candidates présenties : SSN (Semantic Sensor Network) et SAREF (Smart Appliances REFerence) sont deux standards connus dans le domaine de l'Internet des Objets (IdO). Notre recherche s'est limitée à ces deux ontologies car elles sont largement adoptées par la communauté, disponibles en ligne, avec un développement mature et elles sont maintenues au sein d'institutions de normalisation.

L'article est organisé de la façon suivante. La section 2 présente une brève description du système d'irrigation intelligent et de son cycle de fonctionnement. Ainsi les lecteurs auront une vue d'ensemble de notre système contextuel adaptatif. Ensuite, dans la section 3, nous présentons la méthode d'irrigation utilisée dans notre cas d'usage. La section 4 présente l'inventaire des besoins de modélisation ontologique. Les ontologies considérées pour cette étude sont brièvement présentées dans la section 5. La section 6 analyse en quoi les ontologies candidates répondent aux besoins. Enfin, d'autres recherches dans ce domaine sont présentées dans la section 7. La section 8 conclue ce travail et présente des perspectives.

## 2 Définitions des systèmes d'irrigation intelligents

Nous proposons d'adopter les définitions suivantes:

- "Un système contextuel est un système qui utilise le contexte pour fournir des informations et des services appropriés à l'utilisateur. Il convient de noter que la pertinence d'une information ou d'un service dépend de la tâche réalisée par l'utilisateur." (Abowd *et al.*, 1999)

- "Un système contextuel adaptatif est un système contextuel capable de modifier son comportement en fonction des changements du contexte de l'application" (Efstratiou, 2004). Par exemple, un système de gestion des inondations en charge de la surveillance d'un bassin versant est un système contextuel adaptatif s'il envoie des alertes à ses utilisateurs pour les informer des risques de crues et s'il modifie la fréquence de communication de ses noeuds en fonction de ces risques (Sun *et al.*, 2016).

Le contexte dans ce type de système est défini comme "l'ensemble des informations utilisées pour caractériser la situation d'une entité. Une entité peut être une personne, un lieu ou un objet jugé pertinent dans les interactions entre l'utilisateur et l'application"(Abowd *et al.*, 1999). Deux types de contexte sont définis (Sun, 2017):

- Le contexte de bas niveau contient des données quantitatives telles que les mesures issues de capteurs.

- Le contexte de haut niveau, quant à lui, est constitué des données qualitatives qui sont spécifiées en fonction des objectifs de l'application. Un exemple de contexte de haut niveau pour un système d'irrigation automatique est l'état des parcelles agricoles : lorsqu'une parcelle atteint l'état «sol sec», le système déclenche une action d'irrigation.

Un système contextuel adaptatif pour l'irrigation est composé de trois composantes spécifiques: un réseau de capteurs sans fil (RCSF) en charge de la surveillance de l'environnement; un outil d'aide à la décision (OAD) pour envoyer des notifications aux agriculteurs afin de les aider dans leurs décisions d'irrigation et contrôler un système d'irrigation automatique et enfin le système d'irrigation.

Le cycle de fonctionnement d'un système contextuel se découpe en quatre phases : (1) l'acquisition du contexte ; (2) la modélisation du contexte ; (3) le raisonnement sur le contexte ; et (4) la diffusion du contexte (Perera *et al.*, 2014). Dans un système contextuel adaptatif, une phase supplémentaire est ajoutée pour que le système s'adapte aux changements de

contexte. Par conséquent, le cycle de fonctionnement de ce système comprends cinq phases (Sun, 2017).

La figure 1 présente une illustration du cycle de fonctionnement d'un système contextuel adaptatif dédié à notre cas d'usage sur l'irrigation automatique à Montoldre. Ce cas sera sera présenté plus en détail dans la section 3). Ainsi le cycle de fonctionnement est composé de 5 phases:

- Phase d'acquisition du contexte : au cours de cette phase, le système acquiert des données brutes provenant de diverses sources. La principale source de données est le réseau de capteurs sans fils qui mesure, collecte et transmet des mesures brutes. De plus, des données collectées par les stations de météo locale (Roussey *et al.*, 2014) sont aussi transmises au système.

- Phase de modélisation du contexte : les données brutes sont annotées pour pouvoir être intégrées. Ces données sont organisées dans un modèle pour devenir un contexte de bas niveau. Dans le cas d'usage de Montoldre, nous sélectionons les ontologies SSN et SAREF comme deux candidats pour modéliser le contexte.

- Phase de traitement du contexte : au cours de cette phase, un raisonnement est appliqué sur le contexte de bas niveau afin de déduire un contexte de haut niveau. Pour le raisonnement, un moteur à base de règles peut être utilisé.

- Phase de diffusion du contexte : le contexte de haut niveau est distribué aux composants du système ou à d'autres systèmes. Par exemple, le contexte de haut niveau est une entrée de l' OAD de pilotage de l'irrigation.

- Phase d'exploitation du contexte : dans cette phase, le système exploite le contexte pour prendre une décision et lancer une action comme lancer l'irrigation. Le système peut aussi modifier le comportement de ces composants afin qu'ils s'adaptent aux changements du contexte. Par exemple, pendant une forte pluie, le système demande aux nœuds du réseau de passer en mode veille car le système n'aura pas besoin de nouvelles mesures d'humidité du sol suite à la pluie.

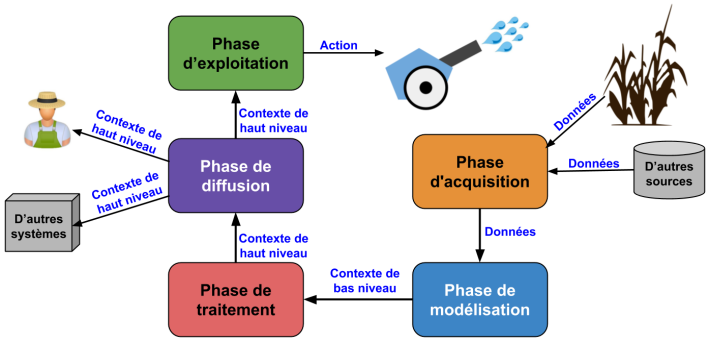



*Figure 1: Le cycle de fonctionnement d'un système d'irrigation intelligent*

## 3 La méthode IRRINOV pour piloter manuellement l'irrigation

Cette section présente la méthode IRRINOV®[1] développée par l'institut technique Arvalis et ses partenaires. Cette méthode propose un guide aux agriculteurs pour prendre des décisions d'irrigation en fonction des mesures des sondes d'humidité du sol et du pluviomètre.

[1]http://www.irrinov.arvalisinstitutduvegetal.fr/irrinov.asp

Autrement dit, la méthode fournit des conseils pour répondre à trois questions : (1) Quand l'irrigation devrait-elle commencer ? C'est à dire quand l'agriculteur doit-il mettre en place son système d'arrosage sur la parcelle (2) Quand lancer un arrosage (démarrage d'un tour d'eau)? (3) Quand l'irrigation devrait-elle s'arrêter ? Autrement dit l'agriculteur peut retirer son système d'arrosage de la parcelle.

La méthode IRRINOV®est constituée d'un ensemble de tables de décision et de recommandations pour gérer l'irrigation d'une parcelle. Cette méthode propose de nombreuses variantes dépendant du type de sol de la parcelle et de sa culture. Nous utiliserons la méthode IRRINOV de la région Limagne pour la culture du maïs grain en sol argilo-calcaire (Arvalis, Limagrain, Chambre d'Agriculture du Puy-de-Dome, 2005).

Les équipements nécessaires pour réaliser des mesures avec la méthode IRRINOV comprennent :

- Une station de mesure IRRINOV®composée de 6 sondes Watermark pour mesurer la tension de l'eau dans le sol (tensiomètre). Parmi les 6 sondes Watermark, 3 sondes sont placées à 30 cm de profondeur dans le sol et les 3 autres sondes sont placées à 60 cm de profondeur.
- Un pluviomètre pour mesurer la quantité d'eau reçue par la culture pendant un tour d'eau.
- Une station météorologique comprenant un pluviomètre pour mesurer la quantité d'eau reçue par la culture pendant une pluie, et un thermomètre pour mesurer chaque jour la température minimale et la température maximale de l'air.

Les sections suivantes décrivent les configurations des équipements et fournissent des recommandations pour avoir un processus d'irrigation de bonne qualité.

### 3.1 Configuration de la localisation des sondes et des équipements

La méthode IRRINOV®spécifie la localisation des équipements de mesure :

- La station IRRINOV doit être située sur le sol dominant de la parcelle à irriguer et elle doit être facilement accessible. La localisation de cette station dépend du système d'arrosage. La station doit être entre deux arroseurs et au moins à 60 mètres du bord de la parcelle. Les sondes doivent être placées sur deux rangs voisins entre des plants comme le montre le schéma de la figure 2.

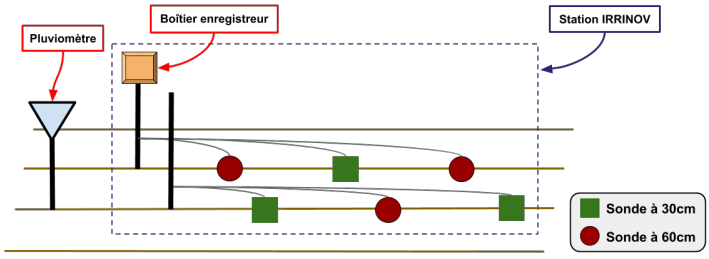



*Figure 2: Plan des sondes Watermark et du pluviomètre dans une parcelle*

- Le pluviomètre mobile doit être proche de la station IRRINOV. Sa hauteur doit être au-dessus de la hauteur maximum de la culture et en dessous de l'arroseur. Les auteurs de la méthode IRRINOV recommandent de placer le pluviomètre sur un pied télescopique pour le maintenir au-dessus de la culture.
- La station météorologique agricole doit être éloignée de tout bâtiment ou arbre et à une hauteur spécifique (inférieure à 2 mètres, soit la hauteur maximale des cultures en champs).

### 3.2 Configuration de la fréquence de mesure

La station IRRINOV®et le pluviomètre mobile doivent être placés dans la parcelle lorsque la culture atteint le stade V2[2]. Les mesures commencent 2 ou 3 jours après l'installation.

Les sondes Watermark doivent fournir une mesure une fois par semaine ou tous les deux ou trois jours si le temps devient sec. De plus, les sondes d'humidité du sol doivent fournir une mesure :

- Avant chaque tour d'eau prévu pour confirmer ou annuler le début d'un nouveau tour d'eau;
- 24 heures à 36 heures après chaque tour d'eau pour évaluer l'efficacité de l'irrigation. Il faut éviter de mesurer l'humidité du sol moins de 24 h après la fin de l'irrigation, car les mesures sont instables.
- Après des pluies pour évaluer leur effet. Par exemple, si la quantité de pluie est inférieure à 10 mm, la date du prochain tour d'eau ne doit pas être modifiée.

L'irrigation doit s'arrêter lorsque la culture atteint le stade R5[3].

### 3.3 Validation des mesures

Pour valider la mesure des sondes Watermark, la méthode IRRINOV définit un écart maximal possible entre les mesures des sondes d'une même niveau de profondeur. Précisément, cet écart est égal à 30 cbar. Si l'écart entre les mesures des sondes est supérieure à 30 cbar, cela signifie que l'une des sondes est hors service et que l'agriculteur doit aller sur le terrain pour recalibrer ou changer la sonde.

Pour obtenir la tension en cbar, la valeur mesurée par la sonde doit être multipliée par un coefficient de correction fourni par le fabricant. Le coefficient de correction est spécifique à chaque lot de sondes : par exemple les sondes de 2003 ont un coefficient de correction égal à 1,7.

Un écart de 10 à 20 cbar de tension entre deux sondes situées à la même profondeur est considéré comme normal. Pour cette raison, la méthode IRRINOV®propose d'installer trois sondes par niveau de profondeur. Les tables de décision d' IRRINOV®proposent des seuils de mesures d'humidité du sol pour lancer l'irrigation ou un tour d'eau. Ce seuil est considéré comme atteint lorsque deux sondes sur trois fournissent une mesure au dessus de ce seuil. Les mesures jugées anormales ne sont pas prises en compte. A noter qu'une valeur mesurée de 199 cbar indique qu'il y a un problème de contact électrique entre la sonde Watermark et le sol.

### 3.4 Présentation des tables de décision

La méthode IRRINOV®(Arvalis, Limagrain, Chambre d'Agriculture du Puy-de-Dome, 2005) propose plusieurs tables de décision pour déterminer le démarrage d'un tour d'eau. Ces tables dépendent du type de sol, de la durée du tour d'eau [4], de la culture et de son stade de développement.

Nous définissons la variable Probe30 (et Probe60) qui représente la mesure atteinte par deux sondes sur les trois sondes situées à 30 cm de profondeur (et respectivement à 60 cm de profondeur).

[2]V2 est un identifiant défini dans (Abendroth *et al.*, 2011) ; il est aussi appelé stade "5 feuilles" conformement à la classification des stades de développement des cultures d'Arvalis

[3]R5 est un identifiant défini dans (Abendroth *et al.*, 2011) et il est aussi appellé stade "grain à 50% humidité" par la classification d'Arvalis

[4]temps entre deux arosages d'une même parcelle

La table de décision 1 définit le seuil pour commencer un tour d'eau pour une culture de mais grain dans un sol argilo-calcaire. Cette table s'applique à la culture de maïs lorsque son stade de développement est entre V2 et V7[5]. Dans cette table, les cellules contiennent le seuil de mesure des sondes mentionné dans la section 3.3 en fonction de la durée des tours d'eau de la parcelle considérée.

La première colonne d'une table de décision doit se lire comme "si la durée du tour d'eau propre à la parcelle est entre 9 à 10 jours" et "lorsque deux sondes à 30 cm de profondeur sur trois (Probe30) sont supérieures à la valeur à 30 cbar et que deux sondes à 60 cm de profondeur sur trois (Probe60) ont une valeur supérieur à 10 cbar " ou " lorsque le total (Probe30 + Probe60) est supérieur à 40 cbar ", donc l'irrigation doit commencer. Il convient à noter que pour cette table de décision, les deux premières lignes sont redondantes, car si $Probe30 > 30$ et $Probe60 > 10$ alors $Probe30 + Probe60 > 40$.

*Table 1: Valeurs de seuil en fonction de la durée des tours d'eau*

| | 9 à 10 jours | 6 à 8 jours | inférieure ou égale à 5 jours |
|---|---|---|---|
| $Probe30$ | 30 cbar | 50 cbar | 60 cbar |
| $Probe60$ | 10 cbar | 20 cbar | 20 cbar |
| total | 40 cbar | 70 cbar | 80 cbar |

### 3.5 Système d'irrigation automatique

À partir de cette méthode de décision dédiée à la prise de décision humaine, nous souhaitons développer un système contextuel adaptatif pour automatiser l'irrigation sur le site de l'AgroTechnoPôle d'Irstea. L'AgroTechnoPôle contient une ferme expérimentale située à Montoldre où les chercheurs peuvent tester leurs prototypes tels que des robots, des équipements de mesure, des machines agricoles et des réseaux de capteurs sans fil. La figure 3 présente les différentes parcelles cultivées appartenant à la ferme en 2018. Une station météorologique Davis Pro 2 est située sur le site. Plusieurs noeuds capteurs sont déployés dans les parcelles pour surveiller l'humidité et la température du sol.

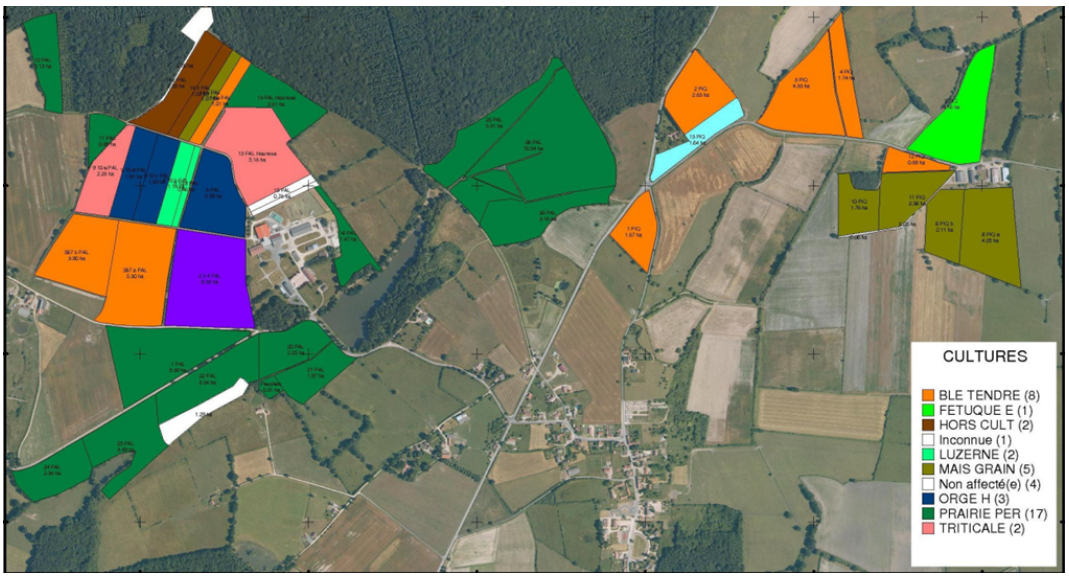



*Figure 3: Parcellaire de l'AgroTechnopole de Montoldre*

[5]V7 est un identifiant défini dans (Abendroth *et al.*, 2011) et il est aussi appelé "10 feuilles" par la classification Arvalis

Le système contextuel adaptatif que nous souhaitons développé doit comporter un réseau de capteurs sans fil contenant des noeuds tensiomètres capable de mesurer l'humidité du sol et des noeuds météo capable de mesurer les pluies.

La localisation des équipements de mesure mentionnées dans la section 3.1, nous précise que la localisation des noeuds du réseau est nécessaire et devra être acquise pendant la phase d'acquisition du contexte.

Les fréquences de mesures présentées dans la section 3.2 définissent les fréquences de mesures et de communication des noeuds du réseau. La détection des événements de pluie ou la détection de la fin des tours d'eau déclencheront des actions de mesure des noeuds. Par conséquent les fréquences de mesure et de communication des noeuds devront s'adapter au contexte.

Les recommandations proposées dans la section 3.3 seront traduites par des règles pour valider les données acquises pendant la phase d'acquisition du contexte. Les tables de décision présentées dans la section 3.4 seront traduites par un ensemble de règles de la forme suivante

```
If (8 < DureeTourEau <11) et (StadeCulture < V7) et ((Probe30 + Probe60) >= 40)
Then (EtatIrrigation = TRUE);
```

Un stade de culture est représenté par une donnée qualitative définie dans un référentiel ordonné comme la classification d'Arvalis. Il faudra développer une fonction correspondant à l'opérateur "<" sur cet ensemble de données qualitatives.

Ces règles font partie de la phase de traitement du contexte pour déduire le contexte de haut niveau à partir du contexte de bas niveau.

## 4 L'extraction des besoins ontologiques

L'objectif de cette section est de présenter les besoins ontologiques de notre système contextuel adaptatif. Pour ce faire, nous avons étudié les données d'entrées nécessaire à la méthode IRRINOV ainsi que ses données de sortie. De plus, nous avons analysé les informations nécessaires pour évaluer la qualité d'un processus d'irrigation. Pour compléter cette étude dédiée à l'irrigation, nous avons utilisé notre expertise sur le déploiement de réseau de capteurs sans fils en milieu ouvert. Nous avions déjà mené plusieurs expérimentations de ce type au sein de l'AgroTechnoPole. Les besoins sont détaillés dans les sections suivantes:

**R1. Déploiement :** Le réseau de capteurs en champs est déployé chaque année. Si le champs doit être labouré il faut retirer les sondes et les noeuds avant de passer les machines. Par conséquent, l'ontologie doit permettre de stocker les informations de déploiement. Ce besoin se découpe en deux sous-besoins.

- **R1.1. Temporalité du déploiement:** Un déploiement d'un réseau de capteur possède une date de début et une date de fin vu que les noeuds capteurs en champs sont retirés chaque année.
- **R1.2. Localisation du déploiement:** La localisation de chaque noeud capteur doit être précisée. L'objectif est d'identifier sur quelle parcelle le déploiement a eu lieu.

**R2. Parcelle:** D'une manière générale, un déploiement d'un réseau de capteurs en champs implique une ou plusieurs parcelles. Dans le cas de capteurs en champs, une parcelle représente la plateforme sur laquelle est installé les noeuds capteurs. Pour l'irrigation il est nécessaire aussi de connaitre les caractéristiques de cette parcelle, telle que sa géométrie, sa surface et la durée du tour d'eau quand elle est irriguée.

**R3. Configuration du réseau:** un réseau de capteurs, composé de plusieurs noeuds, est configuré pour permettre à chacun de ses noeuds de communiquer avec au moins l'un de ses voisins. L'objectif étant que les mesures effectuées par les noeuds puissent être

transmises jusqu'au serveur capable de traiter le contexte. Tous les noeuds sont identifiés par une adresse unique. Cette configuratoin doit aussi être stockée dans l'ontologie. Ce besoin est divisé en quatre sous besoins.

**R3.1. Topologie du réseau:** La topologie d'un réseau de capteur définit les connexion entre les noeuds du réseau. Deux noeuds sont connectés s'ils peuvent communiquer entre eux. La topologie du réseau doit être stockée dans l'ontologie.

**R3.2. Communication du réseau:** Les protocoles de communication utilisés entre les nœuds du réseau doivent aussi être enregistrés.

**R3.3. Status du noeud:** Un noeud peut avoir des status différents: il peut être actif, inactif ou en veille. Le statut d'un nœud à un moment donné doit aussi être conservé. Cette information est importante pour expliquer par exemple des erreurs ou des délais importants lors de la communication entre les noeuds.

**R3.4. Rôle du noeud:** Un noeud dans un réseau possède un rôle donné indiquant les actions de communication dont il est capable. Il peut être un hôte capable de recevoir ou d'envoyer un message, un routeur capable de recevoir plusieurs messages et de les transmettre ou une passerelle capable d'échanger des messages utilisant des protocoles de communication différents.

**R3.5. Localisation du noeud:** la location précise des noeuds doit être conservée. Ce besoin est plus précis que R1.2. Dans le cas de la méthode IRRINOV, la localisation des noeuds est dépendante de la manière dont la culture est implantée. De plus il est nécessaire d'avoir les coordonnées géographiques (latitude, longitude) mais aussi la profondeur dans le sol dans le cas des tensiomètres ou la hauteur dans le cas des noeuds météo. Cette localisation précise permet de calculer la distance entre les noeuds pour évaluer la porter de communication d'un noeud.

**R4. Equipement:** L'ensemble des équipements informatiques, de mesures ou agricoles sont à décrire. Dans le cas de notre système contextuel adaptatif nous pouvons identifier au moins deux catégories d'équipements informatiques : les noeuds capteurs en charge des mesures, les noeuds actionneurs capable de commander un équipement. Nous allons lister de manière exhaustive la liste des équipements nécessaires à notre système contextuel adaptatif dédié à l'irrigation automatique d'une parcelle :

- Noeuds capteurs
  - Des noeuds possédant des sondes Watermark
  - Une station météorologique agricole avec un pluviomètre et un thermomètre
  - Un noeud pluviomètre
- Noeuds actionneurs
  - Les arroseurs
- Noeuds routeurs: noeud en charge de la collecte et de la transmission des mesures brutes fournies par les noeuds capteurs. ils sont aussi en charge de la transmission des commandes jusqu'aux noeuds actionneurs.
- Serveur: équipement informatique sur lequel est installé l'OAD capable de déduire les besoins en irrigation à partir des mesures des tensiomètres et du pluviomètre.

Ainsi nous pouvons spécifier quatre nouveaux besoins:

**R4.1. Capteur:** la description des noeuds capteurs doit être présente dans l'ontologie.

**R4.2. Actionneur:** les noeuds actionneurs doivent être décrits dans l'ontologie.

**R4.3. Composition de l'équipement:** La composition des équipements informatiques doit être représentée dans l'ontologie.

**R4.4. Equipement spécifique:** Les équipements agricoles comme le système d'irrigation doivent aussi être décrit dans l'ontologie.

**R5. Mesure:** Les mesures effectuées par les noeuds capteurs sont aussi décrites dans l'ontologie. La description doit contenir: la valeur mesurée, son unité de mesure et la date à laquelle la mesure a été effectuée, la durée de la mesure s'il s'agit d'une mesure sur un intervalle de temps. Par exemple une mesure de température est une mesure instantanée, une mesure de pluie est une mesure sur un intervalle de temps. Pour notre cas d'utilisation, nous pouvons lister les unités de mesure suivantes :

- Degrés Celsius ($^o$C) pour les mesures de température de l'air
- Millimètre (mm) pour les mesures de qualtités d'eau (pluie ou irrigation)
- Centibar (cbar) pour les tensions du sol
- Unité de mesure brute des sondes Watermark : la mesure de l'humidité du sol effectuée par les sondes Watermark. Cette valeur est ensuite transformée en cbar lorsqu'elle est multipliée par le coefficient correcteur.
- Le référentiel utilisé pour identifier les stades de développement de la culture

**R5.1. Unité de mesure spécifique** Lorsque les mesures ne sont pas associées à des unités traditionnelles, il sera nécessaire de décrire ces nouvelles unités de mesures.

**R6. Phénomène observé:** La localisation des noeuds capteurs précise quel phénomène ils sont capables d'observer. Par exemple, un thermomètre situé sur un mur extérieur est dédié à l'observation de l'air, un thermomètre enterré dans la terre est dédié à l'observation du sol. Pour notre cas d'usage, les phénomènes observés sont l'air, les précipitations, l'irrigation, le sol et la culture. La méthode IRRINOV demande à ce que certains phénomènes soient précisés.

**R6.1. Phénomène observé spécifique:** Dans le cas des sondes Watermark de la méthode IRRINOV, le phénomène ne se limite pas à l'observation du sol, mais à l'observation du sol à une profondeur donnée. Par conséquent, il est nécessaire de pouvoir indiquer cette profondeur de sol dans l'ontologie. Si aucune ontologie ne décrit déjà ces phénomènes naturels il faudra pouvoir les définir et les décrire.

**R7. Propriété:** Les propriétés des phénomènes mesurées par les capteurs sont à représenter dans l'ontologie. La liste de propriétés suivantes correspond au cas d'utilisation de Montoldre :

- Humidité du sol à une profondeur donnée,
- Température du sol à une profondeur donnée,
- Quantité d'eau reçue pendant un tour d'eau,
- Quantité de précipitation réçue dans une journée,
- Température maximal et minimal de l'air pendant une journée,
- Stade de développement des plantes.

**R7.1. Propriété spécifique:** Lorsqu'une propriété n'est pas déjà définie dans une ontologie existante, il sera nécessaire de pouvoir définir cette nouvelle propriété propre au cas d'usage agricole.

**R8. Action:** Les actions capables d'être commandées à distance doivent être indiquées.

**R8.1. Action spécifique:** Pour le cas d'usage de Montoldre, la commande porte sur une action d'irrigation. Cette action doit être complétée par plusieurs paramètres tel que la durée et le débit d'arrosage. La méthode IRRINOV ne fournit aucune indication sur ces deux paramètres.

**R9. Culture:** La culture cultivée sur la parcelle doit être indiquée dans l'ontologie. La méthode IRRINOV demande de spécifier les dates à laquelle la culture a atteint un certain stade de développement. Pour automatiser la détection où la culture a atteint un stade de développement il est nécessaire de connaitre la date de semis et la variété semée.

## 5 Description des ontologies SSN et SAREF

Dans ce projet, nous ne considérons que deux ontologies candidates pour modéliser le contexte d'un système adaptatif dédié à l'irrigation automatique. Nous avons fait le choix de limiter notre étude à ces deux ontologies car elles sont portées par des organisations internationales reconnues dans le domaine et toujours en cours d'évolution. L'ontologie Semantic Sensor Network (SSN) est proposée et maintenue par World Wide Web Consortium[6] (W3C). L'ontologie Smart Appliances REFerence (SAREF) est proposée par l'institut européen des normes de télécommunication[7] (ETSI). Les ontologies SOSA/SSN et SAREF sont présentées en détail dans les sections suivantes.

La première version de SSN a été développée par le groupe de travail de W3C intitulé Semantic Sensor Incubator Group (SSN-XG). Le rapport final de ce groupe a été publié sur le site du W3C le 28 juin 2011. L'objectif de cette ontologie est de permettre à un réseau, ses capteurs et les données associées (mesures) d'être décrits par des modèles de données structurés afin d'en faciliter la gestion, l'interrogation et la compréhension (Compton *et al.*, 2012). Une nouvelle version de cette ontologie a été développée sous l'égide du W3C et de l'Open Geospatial Consortium (OGC) pour intégrer la prise en compte de données spatiales. Cette nouvelle version inclue le patron de conception Sensor, Observation, Sampler et Actuator (SOSA)[8] qui est une évolution du patron de conception Stimulus, Sensor, Observation (SSO)footnote`http://www.w3.org/ns/ssn` de l'ontologie orginale SSN. La nouvelle version de SSN que nous noterons SSN/SOSA devrait être plus simple à utiliser. Elle intégre des nouvelles classes pour représenter les actionneurs et les échantillons. Elle est devenue une recommandation du W3C et de l'OGC en octobre 2017. Comme le montre la figure 4 les principales entités de SSN/SOSA sont les capteurs et leurs mesures, les processus d'échantillonnage et leurs échantillons, les actionneurs et leurs commandes.

Ainsi, pour décrire les mesures, SSN/SOSA propose la classe classe `sosa:Observation`. Une mesure est effectuée par un capteur (`sosa:Sensor`). Une observation est une mesure d'une propriété observable (`sosa:ObservableProperty`) d'un phénomène (`sosa:FeatureOfInterest`).

SSN/SOSA décrit les échantillonnages à l'aide de la classe `sosa:Sampling`. Un échantillonnage est produit par un échantillonneur (`sosa:Sampler`). L'échantillonnage concerne un phénomène (`sosa:FeatureOfInterest`). Les résultats de l'échantillonnage sont des échantillons (`sosa:Sample`).

Enfin, pour décrire les commandes ou actions, SSN/SOSA définit la classe `sosa:Actuation`. Une commande est produite par un actionneur (`sosa:Actuator`). Une commande porte sur une propriété actionnable(`sosa:ActuationProperty`) d'un phénomène ou équipement (`sosa:FeatureOfInterest`).

La figure 4 présente les liens de spécialisation entre l'ancienne version de SSN et sa nouvelle version SSN/SOSA. La classe `ssn:Property` se spécialise en `sosa:ObservableProperty` et `sosa:ActuableProperty`. La classe `ssn:System` se spécialise en `sosa:Actuator`, `sosa:Sensor` et `sosa:Sampler`.

[6] `https://www.w3.org`

[7] `http://www.etsi.org`

[8] `http://www.w3.org/ns/sosa`

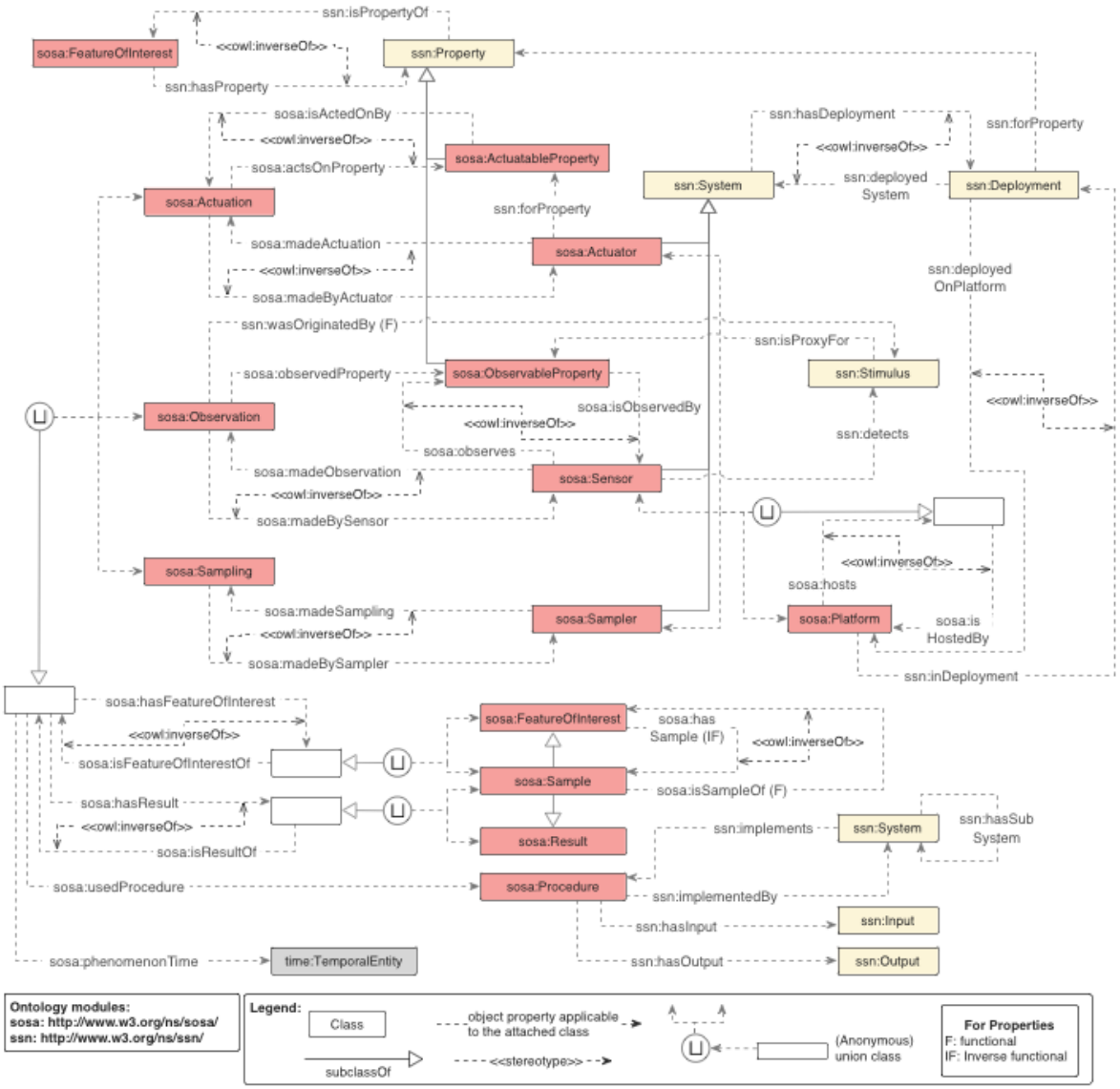



*Figure 4: SSN-SOSA graphical overview*

L'ontologie SAREF[9] est dédiée à la domotique. Son objectif est de faciliter l'intéropérabilité entre objets intelligents et le développement de l'internet des objets. SAREF est une spécification technique de l'ETSI (ETSI, 2015, 2017a). Dans un premier temps, SAREF propose un modèle de base. Ce modèle est ensuite étendu et spécialisé pour couvrir les besoins spécifiques de nouveau domaine. Par exemple il existe une extension spécifique à SAREF pour les bâtiments: SAREF4BLDG (ETSI, 2017b).

La figure 5 présente une version synthétique de l'ontologie SAREF. La classe principale `saref:Device` défini les objets intelligents. Elle se spécialise en capteur (`saref:Sensor`) et en actionneur (`saref:Actuator`). Un objet possède une ou plusieurs fonctions `saref:Fonction`). Une fonction est associée à une ou plusieurs commandes (`saref:Command`). Les commandes changent l'état (`saref:State`) de l'objet. Un objet affiche des services (`saref:Service`) pour rendre ses fonctions accessibles depuis un réseau.

Un objet est aussi décrit par la tâche (`saref:Task`) qu'il peut accomplir pour

[9] `http://w3id.org/saref` `http://saref.linkeddata.es/`

l'utilisateur. Par exemple, un arroseur accomplit une tâche d'irrigation. Un objet intelligent est capable d'effectuer des mesures (`saref:Measurement`). Une mesure a une unité ( `saref:UnitOfMeasure`). Une mesure concerne une propriété ( `saref:Property`).

Enfin, un objet se caractèrise aussi par sa consommation énergétique ( (`saref:Profile`) ou sa consommation d'autres ressources (`saref:Commodity`). Une consommation correspond à une période donnée (`saref:Time`) et a un prix (`saref:Price`).

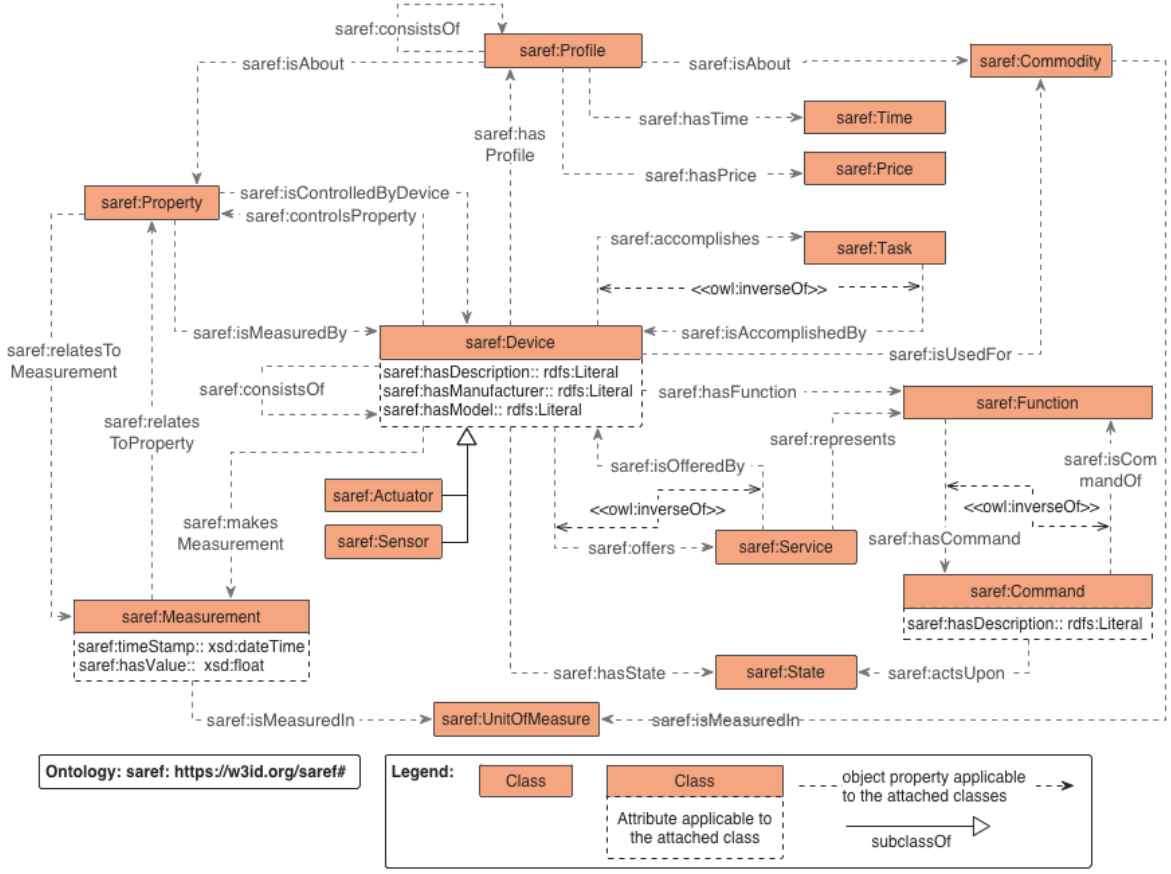



*Figure 5: SAREF graphical overview.*

## 6 Analyse des besoins ontologiques

Dans cette section, une analyse de la couverture des besoins ontologiques par les ontologies SNN/SOSA et SAREF est présentée. Pour se faire, une table de correspondance entre les besoins et les entités définies dans ces ontologies est tout d'abord fournie. Une discussion sur les besoins couverts et sur les questions qui en découlent est ensuite proposée.

### 6.1 Couverture des besoins ontologiques

La table 2 ci-après présente les relations entre les besoins extraits de la section 4 et les entités définies dans les ontologies SSN et SAREF. Dans cette table de correspondance chaque ligne représente un besoin ontologique. Ces besoins sont listés dans la colonne de gauche. Les informations présentent dans les colonnes du milieu et de droite se rapportent respectivement aux ontologies SSN/SOSA et SAREF. Cette table de correspondance s'interprète de la manière suivante : une cellule vide indique qu'un besoin n'est pas couvert par une ontologie

; (b) lorsqu'un besoin est couvert par une ontologie on indique dans la case correspondante les éléments (classes ou propriétés) de l'ontologie permettant de couvrir ce dernier (pour cela on utilise la notation suivante : préfixe de l'ontologie : identifiant de l'entité) ; (c) lorsqu'une note apparaissant sur la documentation fournie avec une ontologie peut être utile pour un besoin (que ce dernier soit couvert ou non), on indique le numéro de cette note entre parenthèse dans la case correspondante.

*Table 2: Couverture des besoins ontologiques par SSN et SAREF.*

| **Requirement** | **SSN/SOSA** | **SAREF** |
|---|---|---|
| R1 Déploiement | ssn:Deployment | |
| R1.1 Temporalité du déploiement | | |
| R1.2 Localisation du déploiement | (1) | |
| R2 Parcelle | sosa:Platform | |
| R3 Configuration réseau | | |
| R3.1 Topologie du réseau | | |
| R3.2 Communication du réseau | | |
| R3.3 Status du noeud | | saref:State |
| R3.4 Role du noeud | | saref:Task |
| R3.5 Localisation du noeud | (1) | |
| R4 Equipement | ssn:System | saref:Device |
| R4.1 Capteur | sosa:Sensor | saref:Sensor |
| R4.2 Actionneur | sosa:Actuator | saref:Actuator |
| R4.3 Composition de l'équipement | ssn:hasSubSystem | saref:consistsOf |
| R4.4 Equipement spécifique | | |
| R5 Mesures | sosa:Observation (2) | saref:Measurement saref:UnitOfMeasure (3) |
| R5.1 Unité de mesures spécifique | | |
| R6 Phénomène observé | sosa:FeatureOfInterest | |
| R6.1 Phénomène observé spécifique | (1) | |
| R7 Propriété | ssn:Property | saref:Property |
| R7.1 Propriété spécifique | | |
| R8 Action | sosa:Procedure | saref:Function saref:Command |
| R8.1 Action spécifique | | |
| R9 Culture | | |

Comme le montre la table 2, certains besoins ne sont pas directement couverts par les ontologies analysées, mais leurs documentations incluent des directives sur la façon de les traiter. En ce qui concerne les besoins R1.2, R3.5 et R6.1, il convient de mentionner que la documentation de SSN/SOSA suggère d'utiliser geoSPARQL (Perry & Herring, 2012) pour modéliser l'information géographique (note (1) dans la table). Il convient également de noter que la représentation correcte de R3.5 et R6.1 est d'une importance particulière car ces informations permettent de valider les mesures nécessaires à la méthode IRRINOV, comme illustré par exemple dans la figure 2.

En ce qui concerne R5, la documentation de SSN/SOSA propose d'utiliser plusieurs ontologies (note (2) dans la table) : "Quantities, Units, Dimensions and data Types" (QUDT) (Hodgson *et al.*, 2014) , "Ontology of units of Measurements" (OM) (Rijgersberg *et al.*, 2013) et "Unified Code for Units of Measure" (UCUM), (Lefrançois & Zimmermann, 2018). La documentation de SAREF mentionne uniquement la deuxième ontologie (note (3) dans le table).

## 6.2 Discussion et questions ouvertes

Dans ce qui suit, les informations extraites de l'analyse de la couverture des besoins ontologiques présentées ci-dessus sont détaillées.

Tout d'abord, il convient de mentionner que les deux ontologies analysées ont été définies comme des ontologies générales couvrant des concepts de haut niveau. Elles doivent être spécialisés et associées à d'autres ontologies pour décrire des cas d'usages spécifiques. En ce sens, nous ne critiquons pas l'absence de couverture des besoins R4.4, R5.1, R7.1 et R8.1 car ils se réfèrent aux connaissances spécifiques du domaine. Cependant, il est important de définir et de prendre en compte ces besoins lors du développement d'extensions ou de

nouvelles ontologies dédié à un cas d'usage agricole. Plus précisément, les besoins présentés dans ce travail représenteraient un bon point de départ pour le développement d'une extension de SAREF pour l'agriculture.

Il faut souligner esnsuite (c'est la principale observation de cette étude), qu'aucune des ontologies analysées ne permet la description de la configuration du réseau (R3), de la topologie (R3.1) et des communications (R3.2).

En ce qui concerne la nécessité de représenter les différents déploiements (R1) dans lesquels les équipements pourraient être impliqués, l'ontologie SSN/SOSA fournit une certaine couverture parce qu'elle prend en compte les déploiements, les parcelles et les composants des équipements. Cependant, bien qu'elle recommande de représenter les informations géographiques à l'aide d'autres ontologies, elle ne traite pas des caractéristiques temporelles des déploiements qui doivent être représentées dans le domaine agricole. SAREF de son côté, ne permet pas la représentation du déploiement. Par conséquent, nous pouvons affirmer qu'aucune des ontologie ne répond complètement au besoin de décrire le déploiement comme une entité spatio-temporel (R1.1 et R1.2).

Concernant les actions (R8), on remarque que les deux ontologies porposent des suggestions pour les modéliser. L'approche de SSN/SOSA est liée aux algorithmes, aux procédures, etc. Elle inclut des informations sur les entrées requises et les sorties générées. L'ontologie SAREF modélise les fonctionnalités des équipements. Elle inclut des informations pratiques telles que les commandes qui peuvent être exécutées (par exemple «ouvrir», «fermer», etc.). Cependant, aucune des deux ontologies ne permet de représenter les informations orientées web telle que « où peut être exécutée cette action » et « où les données peuvent-elles être récupérées ». En d'autres termes aucune des deux ne permet la représentation de services Web ou la description d'objets Web.

La représentation des cultures (R9) mérite quelques explications. Alors que dans la table 2, il est indiqué que ce besoin n'est pas couvert par les ontologies analysées, ce n'est pas exactement le cas pour SSN/SOSA. Une culture peut être modélisée comme un phénomène `sosa:FeatureOfInterest`. Il est envisageable dans un future proche d'imaginer que des capteurs vont pouvoir observer le stade de développement atteint par une culture `ssn:Property`. Dans le cas ou le capteur est positionné sur une plante, cette plante est définit comme étant la plateforme `sosa:Plateform` sur laquelle est positionné le capteur.

En ce qui concerne les questions laissées ouvertes pour la représentation du cas d'usage de Montoldre, il convient de noter que les entités `saref:Profile` et `saref:Commodity` de l'ontologie SAREF pourraient être intéressantes pour représenter la consommation énergétique et la consommation d'eau réalisées par le système d'irrigation intelligent.

Enfin, un autre problème qui n'est pas représenté dans ces ontologies est la validation des mesures effectuées par les capteurs. Par exemple, une mesure de 200 cbar pour une sonde Watermark est considérée comme un indicateur d'un problème de fonctionnement.

## 7 Travaux connexes

Il existe plusieurs systèmes contextuels adaptatifs qui utilisent des ontologies pour intégrer des données hétérogènes et raisonner sur ces données à l'aide de regles. Nous pouvons citer par exemple les travaux de (Goumopoulos *et al.*, 2009) qui présente une ontologie dédiée à l'agriculture de précision. Cette ontologie modélise les caractéristiques des plantes, l'état hydrique de la plante, les paramètres environnementaux, les capteurs et les actionneurs. Ces données sont ensuite utilisé dans un Outil d'Aide à la Décision pour piloter l'irrigation.

Une autre application utilisant des ontologies dans le domaine de la culture des plantes est l'approche détaillée dans (Li *et al.*, 2013). La méthode présentée dans ce travail combine une ontologie du domaine et une ontologie des tâches. L'ontologie du domaine représente le sol, les semences et les machines agricoles tandis que l'ontologie des tâches porte sur les processus entrant en jeux dans la culture des plantes : la sélection du sol, la sélection des graines, la fertilisation et l'irrigation.

Nous pouvons également mentionner l'ontologie présentée dans (Wang *et al.*, 2015) pour la production d'agrumes. Dans ces travaux, les auteurs présentent une ontologie dédiée à la

culture des agrumes. L'ontologie réutilise certains termes de l'ontologie Agricultural Ontology Service (AOS). Elle comprend la modélisation du déséquilibre en nutriments, la modélisation de la fertilisation, de l'irrigation et du drainage des plantes.

Il est à noter que les ontologies mentionnées dans ces articles ne sont pas accessibles en ligne, il n'est donc pas possible de les réutiliser et de vérifier leur contenu.

Le but de la nôtre étude n'est pas de fournir une ontologie pour des systèmes contextuels dédié à l'irrigation ni d'évaluer les ontologies existantes dans le domaine agricole. L'objectif de notre travail est d'analyser deux ontologies standards utilisées dans le domaine de l'Internet des Objets qui pourraient être réutilisées pour développer un système d'irrigation intelligent. Dans ce contexte, nous pouvons mentionner le travail présenté par (Moreira *et al.*, 2017) qui aligne les ontologies SSN et SAREF pour décrire un équipement. Il est à noter que cette analyse se base sur la première version de SSN.

Enfin, les ontologies SSN/SOSA et SAREF ont aussi été alignées dans le travail présenté par (Lefrançois, 2017). Ce dernier présente un alignement entre l'ontologie SEAS et les deux ontologies. Ce travail propose des bonnes pratiques et des patrons pour faciliter la maintenance des ontologies SEAS et SAREF. En résumé, à notre connaissance, il n'existe pas d'étude sur la réutilisation d'ontologies standard pour développer des cas d'usage agricoles.

## 8 Conclusions et perspectives

Cet article a présenté un ensemble de besoins ontologiques pour développer des systèmes contextuels adaptatifs dans le domaine agricole. Ces besoins ont été spécifiés à partir d'un site pilote spécifique et d'une méthodologie d'irrigation donnée. Cet article n'est qu'une étude préliminaire. Ce travail est d'intéret aussi pour le développement de l'extension de SAREF pour le domaine agricole (SAREF4AGRI). Les ontologies SSN et SAREF ont été comparées pour identifier en quoi elles répondaient aux besoins. Certains besoins ne sont pas couverts car ils sont trop spécifiques au cas d'usage considéré. Ces ontologies sont des ontologies de domaine qui doivent être spécialisées pour couvrir ce cas précis.

Il existe cependant des besoins indépendants du cas d'usage qui ne sont couverts par aucune des deux ontologies considérées, par exemple les caractéristiques du réseau.

Pour construire l'ontologie qui permettra de modéliser l'intégraliter du cas d'usage de l'AgroTechnoPole de Montoldre, il faudra combiner plusieurs ontologies. Nous pourrons dans un premier temps aligner les ontologies SSN et SAREF pour évaluer laquelle couvre le plus de besoins. D'autres ontologies auront besoin d'etre associées, comme une ontologie sur la description du réseau ou une ontologie sur la description des services web telle que Web of Things[10] et oneM2M[11].

Enfin, une dernière piste de travail est de trouver l'ensemble des regles à appliquer sur les données du contexte pour automatiser l'irrigation. Rappelons que la méthode IRRINOV est une mdéthoe manuelle basée sur une décision humaine à la lecture de mesures des sondes.

## Acknowledgements

Cette recherche est financée au travers de différents projets : (1) le projet CPER "ConnecSens", (2) le projet "Programme I-Site Clermont WOW! Wide Open to the World - CAP20-25" (16-IDEX-0001) et (3) le projet ETSI STF534 "SAREF extensions". Le projet CPER "ConnecSens" est cofinancé par la région Auvergne-Rhône-Alpes en France, ainsi que par l'Union européenne via le fond FEDER.

## References

ABENDROTH L. J., ELMORE R. W., BOYER M. J. & MARLAY S. K. (2011). Corn growth and development.

[10] https://www.w3.org/TR/wot-thing-description/#vocabularyDefinitionSection

[11] http://www.onem2m.org/technical/onem2m-ontologies

ABOWD G. D., DEY A. K., BROWN P. J., DAVIES N., SMITH M. & STEGGLES P. (1999). Towards a Better Understanding of Context and Context-Awareness. In G. GOOS, J. HARTMANIS, J. VAN LEEUWEN & H.-W. GELLERSEN, Eds., *Handheld and Ubiquitous Computing*, volume 1707, p. 304–307. Springer Berlin Heidelberg.

ARVALIS, LIMAGRAIN, CHAMBRE D'AGRICULTURE DU PUY-DE-DOME (2005). *Guide de l'utilisateur, Carnet de terrain, Pilotez l'irrigation avec la méthode IRRINOV*. Rapport interne, Arvalis.

COMPTON M., BARNAGHI P., BERMUDEZ L., GARCÍA-CASTRO R., CORCHO O., COX S., GRAYBEAL J., HAUSWIRTH M., HENSON C. & HERZOG A. (2012). The SSN ontology of the W3c semantic sensor network incubator group. *Web semantics: science, services and agents on the World Wide Web*, **17**, 25–32.

EFSTRATIOU C. (2004). *Coordinated Adaptation for Adaptive Context-Aware Applications*. Lancaster University.

ETSI (2015). *ETSI TS 264 411 - v1.1.1. SmartM2M; Smart Appliances; Reference Ontology and oneM2M Mapping*. Rapport interne, ETSI.

ETSI (2017a). *ETSI TR 103 411 - v1.1.1. SmartM2M; Smart Appliances; SAREF extension investigation*. Rapport interne, ETSI.

ETSI (2017b). *ETSI TS 103 410-3 - v1.1.1. SmartM2M; Smart Appliances Extension to SAREF; Part3: Building Domain*. Rapport interne, ETSI.

GOUMOPOULOS C., KAMEAS A. D. & CASSELLS A. (2009). An ontology-driven system architecture for precision agriculture applications. *International Journal of Metadata, Semantics and Ontologies*, **4**(1-2), 72–84.

HODGSON R., KELLER P. J., HODGES J. & SPIVAK J. (2014). Qudt-quantities, units, dimensions and data types ontologies. *USA, Available from: http://qudt. org [March 2014]*.

LEFRANÇOIS M. (2017). Planned ETSI SAREF extensions based on the w3c&ogc sosa/ssn-compatible SEAS ontology pattern.

LEFRANÇOIS M. & ZIMMERMANN A. (2018). The Unified Code for Units of Measure in RDF: cdt:ucum and other UCUM Datatypes. *ESWC 2018*.

LI D., KANG L., CHENG X., LI D., JI L., WANG K. & CHEN Y. (2013). An ontology-based knowledge representation and implement method for crop cultivation standard. *Mathematical and Computer Modelling*, **58**(3-4), 466–473.

MOREIRA J. L. R., DANIELE L., PIRES L. F., VAN SINDEREN M., WASIELEWSKA K., SZMEJA P., PAWLOWSKI W., GANZHA M. & PAPRZYCKI M. (2017). Towards iot platforms' integration semantic translations between W3C SSN and ETSI SAREF.

PERERA C., ZASLAVSKY A., CHRISTEN P. & GEORGAKOPOULOS D. (2014). Context aware computing for the internet of things: A survey. *IEEE Communications Surveys & Tutorials*, **16**(1), 414–454.

PERRY M. & HERRING J. (2012). Ogc geosparql-a geographic query language for rdf data. *OGC implementation standard*.

RIJGERSBERG H., VAN ASSEM M. & TOP J. (2013). Ontology of units of measure and related concepts. *Semantic Web*, **4**(1), 3–13.

ROUSSEY C., BERNARD S., ANDRÉ G., CORCHO O., DE SOUSA G., BOFFETY D. & CHANET J.-P. (2014). Weather Station Data Publication at Irstea: an Implementation Report. *Terra Cognita and Semantic Sensor Networks*, p.̃89.

SCHRIJVER R., EUROPEAN PARLIAMENT, EUROPEAN PARLIAMENTARY RESEARCH SERVICE & SCIENTIFIC FORESIGHT UNIT (2016). *Precision agriculture and the future of farming in Europe: scientific foresight study : study*. OCLC: 1015313923.

SUN J. (2017). *Intelligent Flood Adaptive Context-aware System*. PhD thesis.

SUN J., DE SOUSA G., ROUSSEY C., CHANET J.-P., PINET F. & HOU K. M. (2016). A new formalisation for wireless sensor network adaptive context-aware system: Application to an environmental use case. In *Tenth International Conference on Sensor Technologies and Applications SENSORCOMM 2016*, p. 49–55.

UNESCO (2014). The united nations world water development report 2014.

WANG Y., WANG Y., WANG J., YUAN Y. & ZHANG Z. (2015). An ontology-based approach to integration of hilly citrus production knowledge. *Computers and Electronics in Agriculture*, **113**, 24–43.

# Annotation sémantique pour une interrogation experte des Bulletins de Santé du Végétal

Catherine ROUSSEY[1], Tayeb ABDERRAHMANI GHORFI[1]

UR TSCF Irstea, Aubière, France
prenom.nom@irstea.fr

**Résumé** : Le corpus des Bulletins de Santé du Végétal est actuellement disponible sur le Web de données. Les annotations associées à chaque bulletin sont organisées par des propriétés issues du vocabulaire du Dublin Core. Ces annotations répondent à un besoin de recherche documentaire basé sur trois composants : régions françaises, période de publication et types de cultures. Nous avons appliqué une des méthodes de construction d'ontologies de Neon pour faire évoluer le modèle d'annotation des bulletins. Ce modèle contient une ontologie des observations des parcelles et un modèle d'annotation liant le contenu textuel à une entité définie dans l'ontologie précédente. Ces nouveaux modèles répondent à des besoins d'information exprimés par les agronomes.

**Mots-clés** : développement d'ontologies, annotation, Bulletin de Santé du Végétal, competency questions, SPARQL

## 1 Introduction

Lors du projet Vespa, un corpus de bulletins d'alerte agricole intitulé Bulletins de Santé du Végétal (BSV) a été publié sur le Web de données liées (Roussey *et al.*, 2017). Pour ce faire, un modèle d'annotation a été construit en réutilisant des propriétés du vocabulaire du Dublin Core. Ce modèle répond à un besoin de recherche documentaire classique. Un utilisateur va rechercher un ensemble de bulletins en spécifiant une ou plusieurs régions spatiales, une période de publication et un ou plusieurs types de culture.

— Dans ce contexte, une annotation spatiale établit un lien entre un bulletin et une entité géographique. L'entité géographique est définie dans un jeu de données représentant les régions de France avec leurs liens d'inclusion spatiale.
— Une annotation temporelle associe un bulletin à sa date de publication.
— Une annotation thématique établit un lien entre un bulletin et un concept issu d'un thésaurus organisant les types de culture de façon hiérarchique.

Dans le cadre de la recherche documentaire, que nous intitulerons recherche thématique, une annotation établit un lien entre une entité textuelle (le bulletin) et une entité sémantique issue d'une ressource sémantique (le concept d'un thésaurus). La ressource sémantique, d'où est extraite l'entité sémantique, ne possède qu'un seul type d'entité et les organise à l'aide d'une relation hiérarchique représentant soit une inclusion spatiale (entre *ancienne* et *nouvelle* région), soit une relation de généricité entre thèmes (par exemple une culture de céréales est plus générique qu'une culture de blé). Cette ressource est alors intitulée « système d'organisation des connaissances ». Un thésaurus est un bon exemple de ce type de ressource.

À la fin du projet Vespa, de nouveaux besoins ont été exprimés par des agronomes. Ils souhaitaient être capables de retrouver les bulletins à partir des observations réalisées sur les parcelles cultivées. Ces observations sont référencées dans les bulletins. Ce nouveau type de recherche, que nous intitulons recherche experte, implique :

— de proposer un modèle pour décrire ces observations et les parcelles associées,
— d'être capable d'extraire des bulletins les entités représentant ces observations,
— de proposer un modèle d'annotation pour lier le contenu du bulletin aux observations,
— de vérifier que ces nouvelles annotations ne sont pas en contradiction avec les annotations concernant la recherche documentaire thématique. Il serait en effet étonnant de retrouver des observations concernant des parcelles de blé dans un bulletin annoté précédemment avec le concept "vigne".

Ce besoin de recherche experte revient à développer une nouvelle ontologie décrivant les observations et les parcelles. Cette ontologie devra intégrer les systèmes d'organisation des connaissances (le thésaurus des cultures et le jeu de données des régions) utilisés pour la recherche thématique.

Pour ce faire, nous avons choisi d'utiliser une méthode de développement d'ontologie travaillant avec des « *competency questions* » (CQ). Les *competency questions* sont des questions en langue naturelle utilisées pour spécifier les besoins lors de la construction d'une ontologie à partir de zéro (Suárez-Figueroa *et al.*, 2009). Ces questions sont ensuite traduites en langage formel pour construire l'ontologie (Grüninger & Fox, 1995). Plusieurs méthodes utilisent les CQ pour construire une ontologie. Nous pouvons entre autres citer la méthode *Extrem Design*, utilisée pour construire des patrons de conception ontologique (Presutti *et al.*, 2009). Dans notre cas nous avons utilisé les CQ pour développer l'ontologie décrivant les observations et modifier le modèle d'annotation des BSV existant. Nous avons appliqué la méthode de conception d'ontologie proposée dans le scénario 3 de la méthodologie Neon (Suárez-Figueroa *et al.*, 2012).

Cet article est structuré de la manière suivante : La première section présente le corpus annoté et son modèle d'annotation existant. Ensuite, le nouveau besoin d'informations et les CQ associées sont décrits dans la section 3. La nouvelle ontologie et le modèle d'annotation associé sont ensuite décrits dans la section 4. La validation de l'ontologie est décrite dans la section 5. Puis nous concluons avec les perspectives qui émergent de nos travaux.

## 2 Contexte

Nous présentons dans cette section le corpus des Bulletins de Santé du Végétal ainsi que le modèle d'annotation existant pour la recherche de bulletins. Le corpus est actuellement publié sur le Web de données à l'adresse `http://ontology.irstea.fr/bsv/snorql/`

### 2.1 Bulletin de Santé du Végétal

En France, le Grenelle de l'environnement et le plan Ecophyto ont renforcé les réseaux nationaux de surveillance sur les cultures et les pratiques agricoles. Les Bulletins de Santé du Végétal sont une des modalités mises en place par ces réseaux de surveillance dans l'ensemble des régions et départements d'outre-mer. Le Bulletin de Santé du Végétal (BSV) est un document d'information à la fois technique et réglementaire, rédigé sous la responsabilité d'un comité régional d'épidémiosurveillance. Le BSV a pour objectif de réunir et présenter les actualités majeures concernant l'état sanitaire des cultures. Il repose d'un côté sur des analyses du risque phytosanitaire à venir et d'un autre sur la diffusion des informations à caractère réglementaire (arrêtés de lutte obligatoire, notes nationales, évolutions de la réglementation, ...) et non réglementaire (éléments de description de la biologie des bioagresseurs ou des méthodes prophylactiques telles que la gestion des intercultures, du travail du sol, du choix des variétés, ...). Afin de mieux distinguer l'expertise de la préconisation, il n'a pas vocation à faire des préconisations d'utilisation de produits phytosanitaires. La figure 1 présente un exemple de première page d'un BSV de la région Bourgogne.

Les BSV sont une synthèse interprétée des observations effectuées en amont sur les cultures par différents organismes collecteurs, des éléments issus des modèles épidémiologiques, de données météorologiques et parfois d'analyse biologique. Les auteurs des BSV décident, lors de leur réunion éditoriale, si une observation doit être considérée comme un phénomène unique localisé ou bien comme relevant d'un phénomène d'ampleur potentiellement importante et suffisamment représentatif pour être signalé. Étant donné que de nombreux problèmes sanitaires sont d'autant mieux gérables qu'ils sont pris précocement, l'exercice s'avère souvent délicat. Ainsi, les BSV ne sont pas une agrégation automatique de données mesurées mais bien une synthèse humaine la plus consensuelle des jugements sur des observations.

Les BSV sont gratuitement accessibles au format pdf sur les sites internet des Chambres Régionales d'Agriculture et des Directions Régionales de l'Alimentation de l'Agriculture et


FIGURE *1 – Un exemple de BSV de la région Bourgogne, rubrique "grande culture", daté du 5 avril 2011*

de la Foret (DRAAF).

Les BSV sont tout d'abord lus par les conseillers agricoles des coopératives et les agriculteurs pour déterminer leurs futures actions sur leurs cultures ou évaluer l'état de leurs cultures par rapport à l'état des parcelles du réseau. Mais ce corpus intéresse aussi les chercheurs en agronomie. Il constitue une archive des événements sanitaires importants perçus sur les cultures au cours du temps.

## 2.2 Modèle d'annotation existant des BSV

Le modèle d'annotation utilisé lors du projet Vespa pour publier les BSV sur le Web est inspiré du modèle d'annotation de la Bibliothéque Nationale de France (BNF) (Lapôtre, 2017). Ce modèle réutilise le vocabulaire du Dublin Core (dcterms) (Weibel *et al.*, 1998). Le Dublin Core propose un ensemble de propriétés pour enregistrer les métadonnées d'une ressource (titre, auteur, format, *etc.*). Le modèle de la BNF dissocie l'œuvre (Les misérables, de Victor Hugo), de son expression (une édition en 10 volumes publiée en 1862), et de sa manifestation physique (le scan de cette édition disponible sur Gallica). Il nous a paru intéressant de conserver la distinction entre l'expression et la manifestation physique, car un même bulletin peut être disponible sur plusieurs sites Web. Pour ce faire, nous avons repris la notion d'objet d'informations proposé dans l'ontologie fondationnelle *Dolce Ultra Light* (DUL) [1]. Cette ontologie est souvent utilisée pour définir des patrons de conception ontologique.

Avant de présenter en détail le modèle d'annotation correspondant à une recherche thématique, nous allons tout d'abord présenter les deux systèmes d'organisation des connaissances que nous avons utilisés.

1. `http://www.ontologydesignpatterns.org/ont/dul/DUL.owl`

— Un thésaurus intitulé FrenchCropUsage [2] a été défini pendant le projet Vespa pour organiser les cultures en fonction de leur destination (alimentation humaine directe, alimentation animale, industrie alimentaire, etc) et du type de système de culture (grande culture, maraîchage, *etc.*...). Ce thesaurus représente le point de vue français. Il est basé sur les définitions du Larousse Agricole et du wikipedia agricole français. Il contient 272 concepts, la profondeur maximale de la hiérarchie est de 6 niveaux. Ce thésaurus est publié sur le Web de données à l'aide du modèle SKOS à l'adresse `http://ontology.irstea.fr/cropusage/`.
— Pour identifier les régions de France nous avons dû aussi créer un jeu de données propre au projet Vespa, car à la date du projet aucune source ne décrivait les anciennes et les nouvelles régions de France. Dans ce jeu de données, chaque région est une instance d'une classe `irstea:Region`. Ces instances sont liées à des instances similaires issues des jeux de l'IGN, de l'INSEE ou de DBpedia.

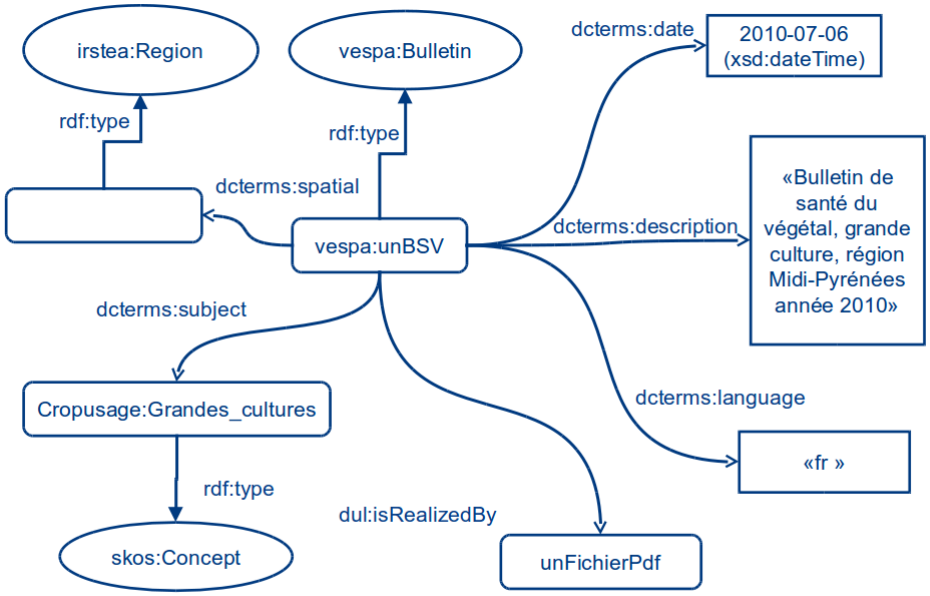



FIGURE 2 – *une exemple d'utilisation du modèle d'annotation décrivant le BSV Grande Culture de la région Midi-Pyrénées du 6 juillet 2010*

La figure 2 représente un exemple d'instanciation du modèle d'annotation thématique des BSV. Les entités que nous avons créées spécifiquement pour l'annotation des BSV sont préfixées par vespa.

Un bulletin est représenté par une instance de la classe `vespa:Bulletin`. Cette classe est une sous-classe de `dul:InformationObject`. Un objet d'information est une entité abstraite qui regroupe l'ensemble des informations relatives à un objet indépendamment de sa matérialisation. Par exemple, l'œuvre "les Misérables" de Victor Hugo est un objet d'information. Les informations de cet objet sont indépendantes d'un exemplaire en particulier. Nous retrouvons cette notion dans la classe `Oeuvre` du modèle d'annotation de la BNF. Un objet d'information peut avoir plusieurs réalisations concrètes distinctes : un fichier pdf, une page html etc. Le lien entre l'objet d'information et sa réalisation (le fichier pdf) est indiqué par la propriété `dul:isRealizedBy`. Les annotations sont portées par l'instance de

2. DOI : 10.25504/FAIRsharing.9228fv et accessible à partir d'agroportal `http://agroportal.lirmm.fr/ontologies/CROPUSAGE`

la classe `vespa:Bulletin`. Les propriétés utilisées pour décrire les BSV sont :

**dcterms :date** : contient la date de publication du bulletin, au format xsd :datetime. Dans le cas d'un bulletin mensuel ou annuel, la date est celle du premier jour de la période.

**dcterms :description** : contient une description textuelle du BSV. Cette description correspond aux informations que nous avons pu extraire des sites Web où le bulletin a été téléchargé : la région correspondant au site Web, l'année de publication et un type de culture qui correspond aux rubriques du site Web où le bulletin apparaît.

**dul :isRealizedBy** est le lien vers le fichier pdf associé.

**dcterms :spatial** est le lien vers le nœud rdf représentant la région dans le jeu de données. Dans l'exemple de la figure 2 ce nœud est intitulé `irstea:places/73` et représente l'ancienne région Midi-Pyrénées.

**dcterms :subject** est le lien vers le `skos:Concept` du thésaurus FrenchCropUsage. Cette propriété peut être utilisée plusieurs fois car un bulletin peut faire référence à différentes cultures.

**dcterms :language** : est la propriété qui stocke la langue du bulletin, dans notre cas uniquement le français (fr).

## 3 Besoins d'informations

Le projet Vespa pour "Valeur et optimisation des dispositifs d'épidémiosurveillance dans une stratégie durable de protection des cultures" avait pour but d'étudier les différentes formes de contribution de l'épidémiosurveillance à la santé des cultures. Lors de ce projet, nous avons construit l'archive des BSV publiée sur le Web présentée dans la section précédente (Roussey *et al.*, 2017). La surveillance des cultures consiste à répertorier l'apparition de bioagresseurs sur les cultures au cours d'une année culturale (de septembre à août). Les bioagresseurs sont les ennemis des cultures, aussi appelés nuisibles des cultures. Ce sont des organismes vivants qui attaquent les plantes cultivées et sont susceptibles de causer des pertes économiques. Ils ont différentes formes :

— Les ravageurs des cultures sont des animaux qui attaquent les plantes cultivées ou les récoltes stockées, en causant un préjudice économique aux agriculteurs.
— Les maladies des plantes empêchent le développement correct des plantes. Elles ont plusieurs causes. Elles peuvent être dues à des champignons parasitaires microscopiques, des bactéries transmises par des insectes suceurs de sève ou par des nématodes dans le sol.
— Les adventices sont les "mauvaises" herbes qui concurrencent les plantes cultivées.

Évaluer le niveau de dangerosité d'un bioagresseur ne se limite pas à observer sa présence sur une parcelle cultivée. En effet, il faut que la plante cultivée ait atteint un certain stade de développement pour que le bioagresseur ait un impact sur la production finale de la parcelle. Il faut aussi parfois que les bioagresseurs soient suffisamment nombreux sur une parcelle pour que les plantes cultivées soient gênées dans leur développement. Lorsqu'un bioagresseur est observé sur une parcelle il faut donc aussi évaluer le niveau de sévérité (sa dangerosité) pour considérer sa présence comme étant une attaque.
Suite au développement de l'archive des BSV et des outils de recherche d'informations mis en place pour interroger ce corpus, les agronomes ont exprimé de nouveaux besoins plus évolués que ceux identifiés au départ du projet. À partir des multiples discussions avec des chercheurs en agronomie qui ont eu lieu durant le projet Vespa, nous avons construit, à la fin du projet, une série de CQ répondant à leurs nouveaux besoins sur le corpus des BSV.

1. Quelles sont les cultures observées en France ?
2. Quelles sont les cultures observées dans la région R (AURA par exemple) ?
3. Quel est l'échantillon de parcelles observées de la culture C dans la région R pendant la période P ?
4. Quels sont les bioagresseurs connus de la culture C (Maïs par exemple) ?

5. Quels sont les ravageurs connus de la culture C ?
6. Quels sont les maladies connues de la culture C ?
7. Quels sont les adventices connues de la culture C ?
8. Quels sont les bioagresseurs connus de la culture C dans la région R ?
9. Quelles sont les attaques du bioagresseur B survenues sur des parcelles de la culture C dans la région R pendant la période P ?
10. Quels sont les stades de développement atteints par la culture C dans les parcelles cultivées de la région R pendant la période P ?
11. Quelles sont les attaques du bioagresseur B survenues sur les parcelles de la culture C dans la région R qui ont atteint un niveau de sévérité S pendant la période P ?
12. Quelle est la chronologie et l'intensité des attaques du bioagresseur B sur les parcelles de culture C dans toutes les régions de France ? (carte de France)
13. Quelles sont les parties de texte dans un ensemble de bulletins qui portent sur des attaques du bioagresseur B sur la culture C ?
14. Quelles sont les parties de texte dans un ensemble de bulletins qui portent sur les stades de développement de la culture C ?
15. Quelles sont les parties de texte dans un ensemble de bulletins qui portent sur les échantillons des parcelles observées de la culture C ?

## 4 Une ontologie d'observation des parcelles et le modèle d'annotation associé

À partir de ces CQ nous voulons modéliser l'ensemble des informations demandées. Il est clair que nous avons besoin d'une ontologie d'observation en environnement naturel. Il existe de notre point de vue deux grandes familles d'observations, les observations correspondant à des mesures automatiques de capteurs (comme les stations météo) et des observations humaines. Plusieurs ontologies correspondent à la première famille car c'est un sujet largement travaillé. Nous pouvons entre autres citer *Semantic Sensor Network* (SSN)(Compton *et al.*, 2012), *Smart Appliances REFerence For Environment* (SAREF4ENVI)(ETSI, 2017), *Observations and Measurements* (OM)(Cox, 2011) pour les plus connues. Concernant la seconde famille, nous ne connaissons que *Extensible Observation Ontology* (OBOE) (Madin *et al.*, 2007) dédiée aux observations scientifiques environementales. À noter que les ontologies dédiées à la description des expérimentations ne font pas partie du périmètre de nos besoins.

Les observations des parcelles sont réalisées par des humains dans le cadre des BSV, mais il est tout à fait envisageable dans un futur proche d'imaginer que ces observations soient automatisées et puissent être réalisées par des équipements de mesures spécifiques. C'est pour cette raison que nous avons sélectionné l'ontologie SSN. Pour SSN, un capteur désigne toute entité capable de suivre une méthode d'observation, que ce soit une personne ou un équipement de mesure [3].

SSN est développée sous l'égide du *World Wide Web Consortium* (W3C). Pour compléter l'ontologie SSN, nous avons sélectionné les ontologies proposées par le W3C en préférant celles qui ont atteint le statut de recommandation. Notre objectif est donc de sélectionner un ensemble d'ontologies du W3C qui répondent aux besoins exprimés par les CQ. Le tableau 1 présente la liste des ontologies sélectionnées.

3. `https://www.w3.org/2005/Incubator/ssn/ssnx/ssn#Sensor`

| Nom | Acronyme | Auteur | Référence |
|---|---|---|---|
| Sensor Observation Sampler Actuator [4] | ssn/sosa | W3C OGC | (Armin *et al.*, 2018) |
| Semantic Sensor Network [5] | ssn | W3C | (Compton *et al.*, 2012) |
| Time [6] | time | W3C | (Hobbs & Pan, 2006) |
| GeoSparql [7] | geo | OGC | (Battle & Kolas, 2012) |
| Prov Ontology [8] | prov | W3C | (Lebo *et al.*, 2013) |
| Event [9] | event | C4DM at QMUL | (Raimond & Abdallah, 2007) |
| Web Annotation Data Model [10] | oa | W3C | (Sanderson *et al.*, 2013) |
| Simple Knowledge Organisation System [11] | skos | W3C | |

TABLE *1 – Liste des ontologies réutilisées*

L'ontologie SSN a évolué (Compton *et al.*, 2012). En 2017, une nouvelle version de cette ontologie construite cette fois-ci sous l'égide du W3C et de l'*Open Geospatial Consortium* (OGC) a été acceptée comme recommandation (Armin *et al.*, 2018). Cette version intègre un nouveau patron de conception intitulé *Sensor Observation Sampler Actuator* (SOSA). SSN/SOSA recommande d'utiliser l'ontologie *GeoSparql* (Battle & Kolas, 2012) pour décrire les entités spatiales. A noter que pour le W3C, l'ontologie *Time* (Hobbs & Pan, 2006) est suffisante pour décrire un évènement (une entité localisée dans le temps). Dans le cas d'une alerte agricole, nous avions besoin d'un modèle simple pour décrire un évènement comme une entité spatialisée et temporalisée impliquant des agents. Nous avons sélectionné l'ontologie *Event* (Raimond & Abdallah, 2007) pour sa simplicité et sa couverture de l'ensemble de nos besoins. Mais d'autres ontologies étaient possibles comme l'ontologie *Linking Open Descriptions of Events* (LODE) (Shaw *et al.*, 2009).

Les sections suivantes présentent les modèles permettant de décrire les observations faites sur les cultures, les alertes agricoles et les liens vers les textes des BSV. A ce stade, ces modèles sont en cours de discussion et ne sont pas formalisés en RDF.

### 4.1 Modèle d'observation des stades de développement des cultures

La figure 3 présente un exemple de description d'une observation du stade de développement atteint par un échantillon de parcelles à l'aide des ontologies SSN/SOSA, GeoSparql, Time, SKOS et QUDT. En plus des ces ontologies, nous réutilisons le thésaurus FrenchCropUsage qui définit des types de cultures à l'aide du modèle SKOS. Donc un type de culture est une instance de la classe `skos:Concept`.

Un nouveau système d'organisation des connaissances est nécessaire pour décrire les stades de développement des cultures. Nous avons sélectionné le référentiel BBCH qui est actuellement décrit au sein d'un jeux de données de l'ontologie CROP [12] disponible sur l'Agroportal du LIRMM. Dans notre exemple, un stade de développement est défini comme une instance de la classe `skos:Concept`.

Nous avons besoin aussi de décrire les unités. SSN préconise plusieurs jeux de données et leurs ontologies associées disponibles sur le Web de données : "*Quantities, Units, Dimensions and data Types*" (QUDT) (Hodgson *et al.*, 2014) , "*Ontology of units of Measurements*"

4. https://www.w3.org/TR/vocab-ssn/
5. https://www.w3.org/2005/Incubator/ssn/ssnx/ssn
6. https://www.w3.org/TR/owl-time/
7. http://www.opengeospatial.org/standards/geosparql
8. https://www.w3.org/TR/prov-o/
9. http://motools.sourceforge.net/event/event.html
10. https://www.w3.org/TR/annotation-model/
11. https://www.w3.org/TR/skos-reference/
12. http://www.cropontology.org/

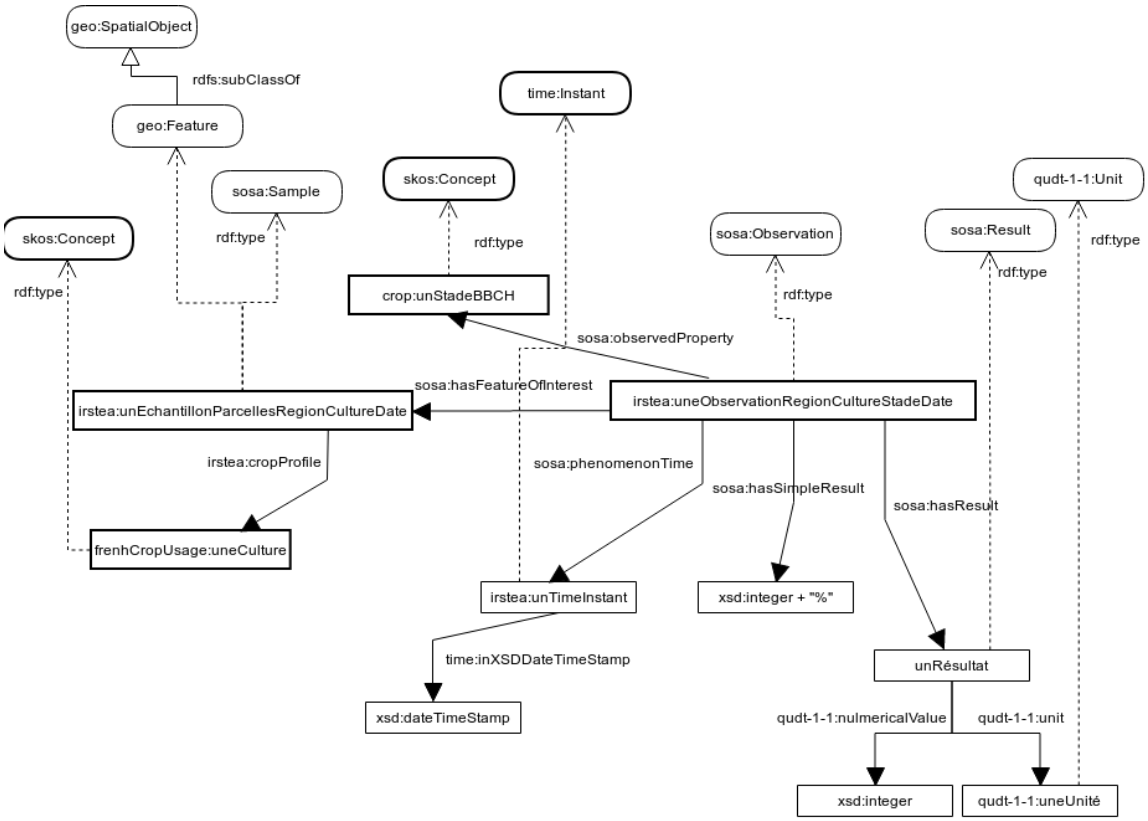



FIGURE *3 – Exemple d'observation de stades de développement*

(OM) (Rijgersberg *et al.*, 2013) et "*Unified Code for Units of Measure*" (UCUM), (Lefrançois & Zimmermann, 2018). Dans notre exemple, nous utilisons la dernière version de l'ontologie QUDT. Ainsi, nous identifions une unité donnée comme une instance de la classe `qudt-1-1:Unit`.

Comme le montre la figure 3, une observation d'un stade de développement d'une culture est identifiée par une instance de la classe `sosa:Observation`. Cette observation porte sur un échantillon de parcelles. Un échantillon de parcelles est un objet géographique. Donc un échantillon est défini comme une instance de `sosa:Sample` et de `geo:Feature`.

Les propriétés utilisées pour décrire l'observation sont :

**sosa :hasFeatureOfInterest** lie une instance de `sosa:Observation` à l'instance représentant l'échantillon de parcelles observées.

**irstea :cropProfile** lie l'instance représentant l'échantillon de parcelles à une instance de `skos:Concept` représentant le type de culture cultivé sur ces parcelles extrait du thésaurus FrenchCropUsage.

**sosa :observedProperty** lie une instance de `sosa:Observation` à une instance de `skos:Concept` représentant le stade de développement observé, extrait du référentiel BBCH décrit dans l'ontologie CROP.

**sosa :phenomenonTime** lie une instance de `sosa:Observation` à une instance de `time:Instant` indiquant la date où cette observation a eu lieu. Dans notre cas il s'agit de la date de publication du BSV. Ainsi, la valeur stockée dans la propriété `time:inXSDDateTimeStamp` doit être la même que celle de `dcterms:date`.

**sosa :hasSimpleResult** est un attribut qui contient un pourcentage. Ce pourcentage indique le ratio entre le nombre de parcelles qui a atteint le stade de développement sur le nombre de parcelles totales composant l'échantillon.

**sosa :hasResult** lie une instance de `sosa:Observation` à une instance de `sosa:Result`. Cette instance possède deux attributs : l'un indique l'unité, l'autre la valeur.

## 4.2 Modèle d'observation d'un bioagresseur dans une culture

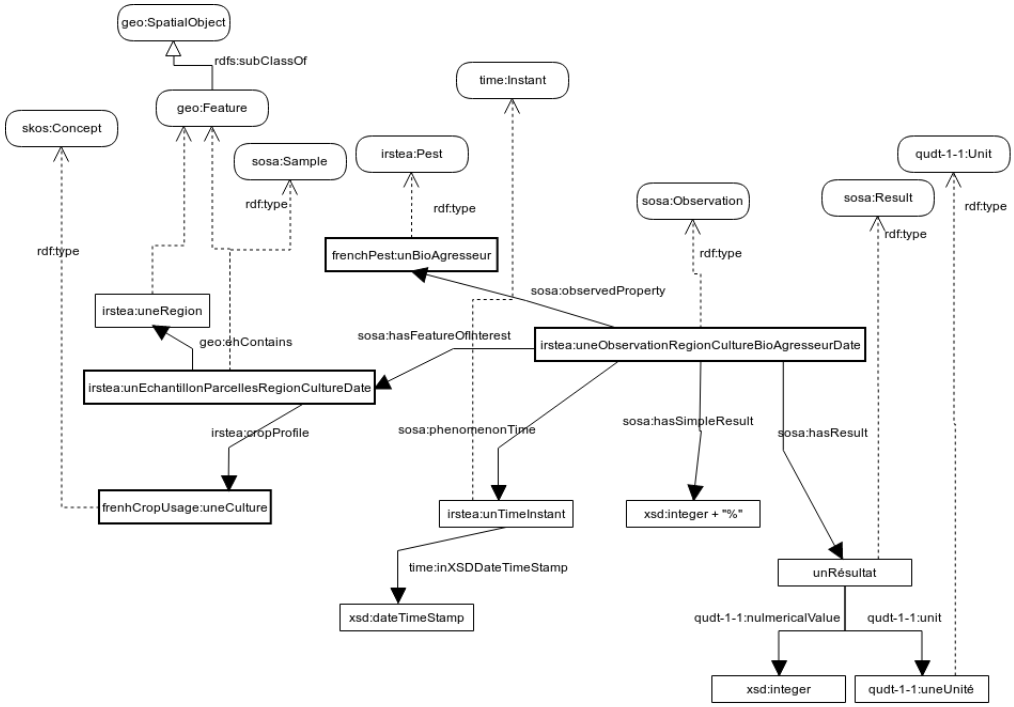



FIGURE *4 – Exemple d'observation de la présence d'un bioagresseur*

La figure 4 présente une description de l'observation de la présence d'un bioagresseur sur un échantillon de parcelles à l'aide des ontologies : SSN/SOSA, Time, GeoSparql, SKOS et QUDT. Nous retrouvons dans cet exemple les références au thésaurus FrenchCropUsage et au jeu de données des unités.

Comme élément nouveau, nous avons indiqué une référence vers le jeu de données décrivant les régions de France. Une région est maintenant une instance de `geo:Feature`. Nous devrons donc faire évoluer le jeu de données existant présenté dans la section précédente (cf figure 2) pour répondre à cette spécification.

Un nouveau jeu de données doit être référencé pour décrire les bioagresseurs des cultures. À ce stade nous n'avons pas encore trouvé de jeu de données préalablement publié sur le Web de données répondant à ce besoin. Il est possible que ce jeu puisse être modélisé en SKOS. Afin de ne pas contraindre la modélisation nous avons représenté un bioagresseur comme une instance de la classe `irstea:Pest` appartenant à un jeu de données intitulé `FrenchPest`.

Une observation de la présence d'un bioagresseur se représente comme une instance de la classe `sosa:Observation`. Les propriétés utilisées pour décrire cette observation sont :

**sosa :hasFeatureOfInterest** lie une instance de `sosa:Observation` à l'instance représentant l'échantillon de parcelles observées.

**irstea :cropProfile** lie l'instance représentant l'échantillon de parcelles à une instance de `skos:Concept` représentant le type de culture cultivé sur ces parcelles.

**geo :ehContains** lie l'instance représentant l'échantillon de parcelles à une instance de `geo:Feature` représentant la région.

**sosa :observedProperty** lie une instance de `sosa:Observation` à une instance de Pest représentant le bioagresseur observé.

**sosa :phenomenonTime** lie une instance de `sosa:Observation` à une instance de `time:Instant` indiquant la date où cette observation a eu lieu. Dans notre cas il s'agit de la date de publication du BSV. Ainsi, la valeur stockée dans la propriété `time:inXSDDateTimeStamp` doit être la même que celle de `dcterms:date`.

**sosa :hasSimpleResult** est un attribut qui contient un pourcentage. Ce pourcentage indique le ratio entre le nombre de parcelles où le bioagresseur a été observé et le nombre de parcelles totales composant l'échantillon.

**sosa :hasResult** lie une instance de `sosa:Observation` à une instance de `sosa:Result`. Cette instance a deux attributs : l'un indique l'unité, l'autre la valeur.

### 4.3 Modèle d'une alerte agricole

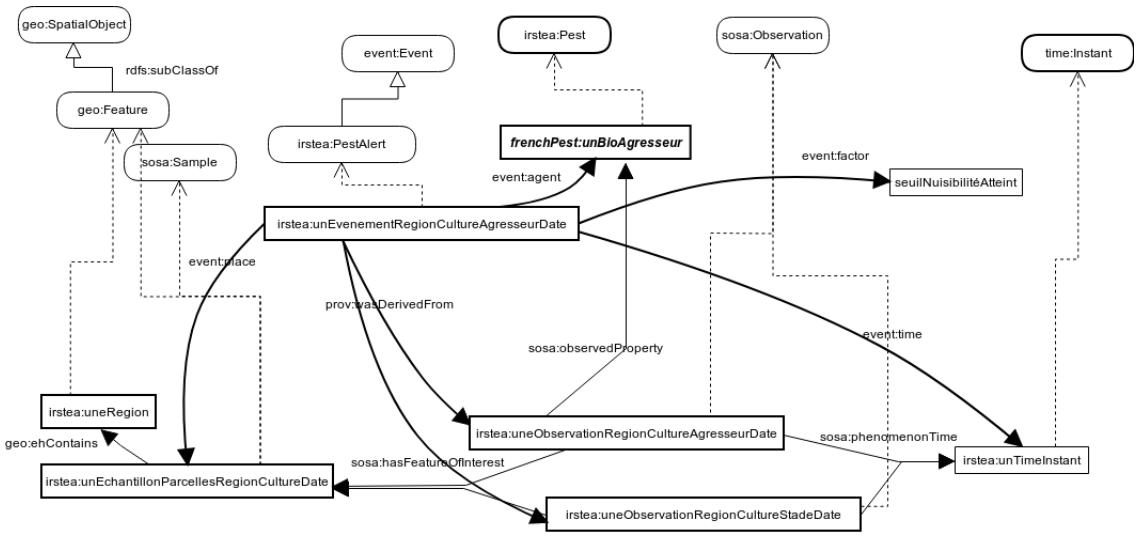



FIGURE *5 – Un exemple de description d'une alerte sur les parcelles*

Une alerte indiquant une attaque d'un bioagresseur sur des parcelles est lancée si plusieurs conditions sont remplies : 1) La culture a atteint un stage de développement donné, 2) La présence du bioagresseur est observée un certain nombre de fois. En effet, certains agresseurs, s'ils arrivent sur la plante lorsqu'elle est jeune ou déjà très mature ne vont pas impacter son développement. De plus, un agresseur peut être présent sur la plante sans provoquer de dégâts suffisants pour impacter le développement de la plante et réduire la production de la parcelle. Par conséquent, la présence du bioagresseur doit être supérieure à un seuil fixé. C'est pourquoi une alerte est liée à l'observation du stade de développement de la plante et aussi à l'observation de la présence d'un bioagresseur.

Pour décrire une alerte nous utilisons les ontologies *Event* et *Prov*. Une alerte est un événement impliquant un lieu, une date, des agents et des facteurs. Dans notre modèle, nous spécialisons la classe `event:Event` pour créer une classe représentant nos alertes agricoles, intitulée `irstea:PestAlert`. À noter que toutes les autres instances sont déjà utilisées dans les modèles d'observation précédents.

Les propriétés utilisées pour décrire une alerte sont :

**event :place** lie une instance de `irstea:PestAlert` à l'instance représentant l'échantillon de parcelles. Cette instance est issue des modèles précédents sur les observations des parcelles.

**event :agent** lie une instance de `irstea:PestAlert` à une instance de Pest représentant le bioagresseur observé.

**event :time** lie une instance de `irstea:PestAlert` à une instance de `time:Instant` pour indiquer la date de l'observation. Cette instance est issue des modèles précédents sur les observations des parcelles.

**event :factor** est un attribut booléen qui indique si le seuil de nuisibilité est atteint.

**prov :wasDerivedFrom** lie une instance de `irstea:PestAlert` aux instances d'observation des stades de développement et de la présence des agresseurs qui ont permis de lancer cette alerte.

### 4.4 Modèle d'annotation des bulletins

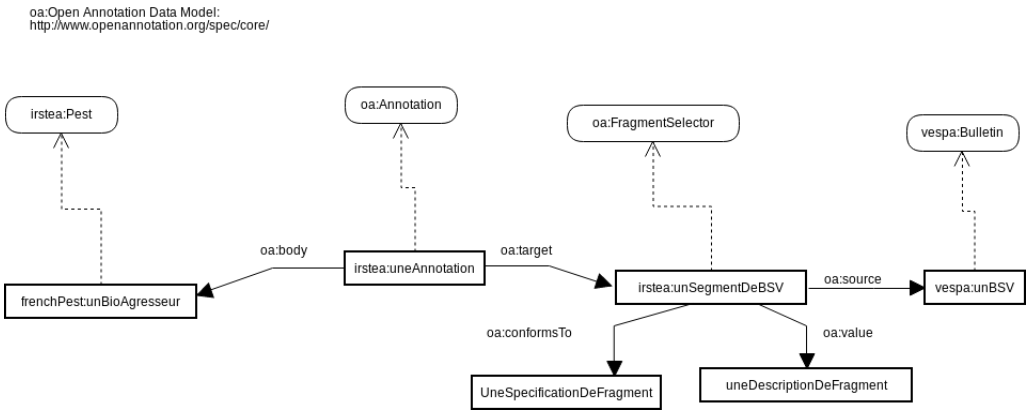



FIGURE *6 – Exemple d'annotation des BSV*

Les figures précédentes décrivent les observations faites sur les parcelles cultivées. La figure 6 décrit le modèle permettant de lier les observations précédentes au texte des BSV. Pour ce faire nous réutilisons l'ontologie "*Web Annotation Data Model*" ou OA du W3C. Les propriétés utilisées pour décrire une instance de `oa:Annotation` sont :

**oa :body** lie une instance de `oa:Annotation` à une instance représentant le sens de l'annotation. Dans la figure 6 il s'agit d'une instance représentant un bioagresseur donné.

**oa :target** lie une instance de `oa:Annotation` à une instance de fragment de texte.

**oa :conformsTo** lie une instance de fragment de texte à un type de sélecteur, le sélecteur étant une fonction qui identifie un fragment de texte. Un fragment de texte peut être identifié par sa position, la chaîne de caractères, etc.

**oa :value** est un attribut qui contient les paramètres du sélecteur pour identifier le fragment.

**oa :source** lie une instance de fragment de texte à son document source. Dans notre cas il s'agit d'une instance de `vespa:Bulletin`.

## 5 Validation

Pour vérifier que nos modèles couvrent bien l'ensemble des besoins exprimés dans les CQ, nous avons demandé à ce qu'un expert en Web sémantique, qui n'a pas participé à la construction des modèles, traduise chacune des CQ en requêtes SPARQL. Cet expert joue le rôle de valideur.

L'ensemble des CQ ont été traduites en requêtes SPARQL. Ces requêtes sont visibles sur le site ontology.irstea.fr à l'adresse `http://ontology.irstea.fr/pmwiki.php/Site/BSVCompetencyQuestions`.

Par exemple, la première CQ sur les cultures observées en France se traduit par la requête :

```
SELECT DISTINCT ?culture
WHERE {
  ?observation a sosa:Observation ;
               sosa:hasFeatureOfInterest ?echantillon .
  ?echantillon irstea:cropProfile ?culture.
}
```

La CQ numéro 9 sur les attaques des bioagresseurs se traduit par :

```
SELECT DISTINCT ?observation
WHERE { ?observation a sosa:Observation ;
                     sosa:hasFeatureOfInterest ?echantillon ;
                     sosa:observedProperty frenchPest:B;
                     sosa:phenomenonTime ?time.
        ?time time:inXSDDateTimeStamp ?stamp.
        ?echantillon geo:ehContains irstea:R ;
                     irstea:cropProfile frenhCropUsage:C.
FILTER ((?stamp > "t1"^^xsd:dateTimeStamp) &&
(?stamp < "t2"^^xsd:dateTimeStamp))
}
```

Le fait d'avoir traduit l'intégralité des CQ laisse supposer que l'ensemble des besoins est couvert par le nouveau modèle d'annotation et les ontologies associées. Une deuxième phase de validation a été réalisée par le concepteur de l'ontologie. Quelques divergences sont apparues entre le concepteur et le valideur lors de ces phases de validation sur les CQ 13 et 14.

### 5.1 Problème de la CQ 13

La CQ numéro 13 porte sur les annotations des attaques des bioagresseurs et elle a été traduite en SPARQL ainsi :

```
SELECT DISTINCT ?text
WHERE { ?observation sosa:hasFeatureOfInterest ?echantillon ;
                     sosa:observedProperty frenchPest:B.
        ?echantillon irstea:cropProfile frenchCropUsage:C.
        ?annotation oa:body frenchPest:B;
                    oa:target ?segment.
        ?segment oa:value ?text.
}
```

Le concepteur espérait que l'annotation lierait l'observation de la présence d'un bioagresseur au texte des BSV. Le valideur a lié le texte à l'instance de bioagresseur. Le concepteur et le valideur ont donc deux interprétations différentes de la même CQ. Chacune de ces interprétations va donner naissancance à une nouvelle CQ et requête SPARQL associée. Ainsi, l'ontologie finale formalisera chacune de ces deux interprétations.

À noter que la différence d'interprétation de la CQ 13 entre le concepteur et le valideur nous montre les limites des modèles. Un modèle (définition de classes et de propriétés) n'est

pas suffisant pour décrire comment utiliser ces classes et ces propriétés. Chacun peut travailler avec les mêmes entités (classe et propriété) mais organisées (liées) de manière différente. Ce qui signifie qu'il faut maintenant clarifier, définir et documenter des "patrons d'usage" pour améliorer la cohérence de jeux de données instanciant les modèles, afin de toujours exprimer la même chose de la même manière.

### 5.2 Problème de la CQ 14

La CQ numéro 14 porte sur les annotations des stades de développement. La première requête SPARQL proposée était la suivante :

```
SELECT DISTINCT ?text
WHERE { ?observation sosa:hasFeatureOfInterest ?echantillon ;
                     sosa:observedProperty ?stage.
        ?stage a skos:concept.
        ?echantillon irstea:cropProfile frenchCropUsage:C.
        ?annotation oa:body ?stage;
                    oa:target ?segment.
        ?segment oa:value ?text.
}
```

Malheureusement, rien dans cette requête n'indique que ce qui est lié au texte est un stade de développement. Il faudrait indiquer le source d'où est extrait le concept représentant le bioagresseur. Une autre solution proposée est de spécialiser la classe observation pour créer une classe "Observation de stade de développement" et une autre classe "Observation de la présence d'un agresseur".

## 6 Travaux connexes

Il existe des modèles pour stocker des observations agricoles. Nous pouvons entre autre citer le modèle *Agricultural Model Intercomparison and Improvement Project* (AGmip) (Porter *et al.*, 2014) basé sur le modèle americano-canadien ICASA. Ces modèles permettent de stocker et d'échanger les résultats des modèles de simulation du développement des cultures. Ce sont des modèles exprimés sous forme tabulaire. Plusieurs ontologies du domaine agricole sont disponibles sur le Web. La plus ancienne est agroRDF, developpée par KTBL (Martini *et al.*, 2013). Cette ontologie est la traduction d'un schema XML agroXML. Elle est monolitique, elle n'intègre pas d'ontologies existantes et elle est faiblement documentée. La plus récente est l'ontologie FOODIE issue du projet européen du même nom (Palma *et al.*, 2016). Elle a pour but de faciliter l'intégration de données issues de fournisseurs différents dans le domaine de l'agriculture de précision. A noter que cette ontologie est en fait la traduction d'un modèle de base de données relationnelle. Par conséquent, les attributs de chaque classe donnent naissance a une nouvelle propriété. Elle ne répond donc pas au principe de modélisation RDFS, où une propriété est une "first class citizen". Autrement dit, une propriété n'appartient à aucune classe. Elle peut s'appliquer à plusieurs classes. Pour limiter la porter d'un propriété il faut définir une contrainte sur cette propriété.

Dans notre cas, nous avons favorisé la réutilisation d'ontologies reconnues sur le Web de données liées pour construire notre nouveau modèle d'annotation des BSV. Notre modèle est devenu un réseau d'ontologies incluant entre autre une ontologie pour les observations, une ontologie pour les événements et le modèle d'annotation du W3C. Un état de l'art de 2006 (Uren *et al.*, 2006) décrivant les fonctionnalités des outils de gestion de connaissances indiquait que peu d'outils d'annotation sont capables de faire évoluer leur modèle d'annotation. A notre connaissance ce constat est toujours d'actualité car faire évoluer le modèle d'annotation implique qu'il faut être capable de faire évoluer les annotations associées. L'évolution des annotations constitue un axe de recherche à part entière (Cardoso *et al.*, 2017).

Les outils d'annotation font l'hypothèse principale qu'une seule ontologie est utilisée pour structurer les annotations. De plus la plupart de ces outils répondent à un besoin de recherche documentaire thématique. Donc les entités sémantiques utilisées dans les annotations sont organisées dans un système d'organisation des connaissances (un thésaurus) et non dans un jeu de données structuré par une autre ontologie. L'évolution des outils d'annotation est de travailler avec plusieurs systèmes d'organisation des connaissances : un par type d'entités de la recherche thématique. Nous pouvons par exemple citer la plateforme KIM (Popov *et al.*, 2004) qui permet de mettre en place plusieurs chaînes de traitements GATE pour reconnaître des entités sémantiques issues de lexiques différents. Cette plateforme utilise une seule ontologie de haut niveau intitulé KIM. Le projet Parmenides a mis en place une autre chaîne de traitements GATE pour extraire des événements structurés par une ontologie dédiée et donc lier des entités sémantiques issues de lexiques différents (Hogenboom *et al.*, 2010).

Dans nos travaux, nous allons développer et instancier un nouveau modèle d'annotation distinct du modèle existant car répondant à des besoins différents. Donc le future modèle ne remplacera pas le modèle existant d'annotation. Par compte, nous devrons vérifier la cohérence des annotations instanciant les deux modèles.

## 7 Conclusion et Perspectives

Le corpus des Bulletins de Santé du Végétal est actuellement disponible sur le Web de données à partir d'un modèle d'annotations basé sur le vocabulaire dublin core. Le modèle existant répond à un besoin de recherche documentaire basé sur trois composants : région spatiale, date de publication et type de cultures.

Nous avons appliqué une méthode de construction d'ontologies de Neon pour faire évoluer le modèle d'annotation des bulletins et définir un modèle d'observation des parcelles. Cette méthode exprime les besoins à partir de competency questions et réutilise des ontologies existantes. Ce nouveau réseau d'ontologies modélise des informations détaillées demandées par les agronomes à la fin du projet Vespa.

Nos travaux futurs porteront tout d'abord sur l'implémentation des modèles présentés dans cet article. Puis, nous peuplerons ces modèles à l'aide des sorties d'une chaîne d'outils de traitement de la langue appliquée sur les BSV. Une fois validés, les résultats devront être publiés sur le Web des données pour compléter la description des BSV existants. Nous devrons aussi travailler sur la cohérence entre les différents modèles et jeux de données associés (l'existant et le futur). Par exemple, nous pourrons vérifier que les échantillons de parcelles culturales explicités dans le modèle d'annotation des observations sont bien inclus dans la région explicitée dans le modèle d'annotation documentaire existant. D'autres types de vérification plus complexes pourront être exploités.

## Références


ARMIN H., KRZYSZTOF J., COX S., LE PHUOC D., TAYLOR K. & LEFRANÇOIS M. (2018). Semantic Sensor Network Ontology.

BATTLE R. & KOLAS D. (2012). Enabling the geospatial semantic web with parliament and geosparql. *Semantic Web*, **3**(4), 355–370.

CARDOSO S. D., REYNAUD-DELAÎTRE C., DA SILVEIRA M. & PRUSKI C. (2017). Combining rules, background knowledge and change patterns to maintain semantic annotations. In *AMIA 2017*.

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

COX S. (2011). Observations and measurements-xml implementation.

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

GRÜNINGER M. & FOX M. S. (1995). Methodology for the design and evaluation of ontologies.

HOBBS J. R. & PAN F. (2006). Time ontology in owl. *W3C working draft*, **27**, 133.

HODGSON R., KELLER P. J., HODGES J. & SPIVAK J. (2014). Qudt-quantities, units, dimensions and data types ontologies. *USA, Available from : http ://qudt. org [March 2014]*.

HOGENBOOM F., HOGENBOOM A., FRASINCAR F., KAYMAK U., VAN DER MEER O., SCHOUTEN K. & VANDIC D. (2010). SPEED : A Semantics-Based Pipeline for Economic Event Detection. In J. PARSONS, M. SAEKI, P. SHOVAL, C. WOO & Y. WAND, Eds., *Conceptual Modeling – ER 2010*, p. 452–457, Berlin, Heidelberg : Springer Berlin Heidelberg.

LAPÔTRE R. (2017). Library metadata on the web : the example of data. bnf. fr. *JLIS. it*, **8**(3), 58.

LEBO T., SAHOO S., MCGUINNESS D., BELHAJJAME K., CHENEY J., CORSAR D., GARIJO D., SOILAND-REYES S., ZEDNIK S. & ZHAO J. (2013). Prov-o : The prov ontology. *W3C recommendation*, **30**.

LEFRANÇOIS M. & ZIMMERMANN A. (2018). The Unified Code for Units of Measure in RDF : cdt :ucum and other UCUM Datatypes. *ESWC 2018*.

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

MARTINI D., SCHMITZ M. & MIETZSCH E. (2013). agrordf as a semantic overlay to agroxml : a general model for enhancing interoperability in agrifood data standards. In *CIGR Conference on Sustainable Agriculture through ICT Innovation*.

PALMA R., REZNIK T., ESBRÍ M., CHARVAT K. & MAZUREK C. (2016). An INSPIRE-Based Vocabulary for the Publication of Agricultural Linked Data. In *Ontology Engineering*, volume 9557, p. 124–133. Cham : Springer International Publishing.

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

PORTER C. H., VILLALOBOS C., HOLZWORTH D., NELSON R., WHITE J. W., ATHANASIADIS I. N., JANSSEN S., RIPOCHE D., CUFI J., RAES D., ZHANG M., KNAPEN R., SAHAJPAL R., BOOTE K. & JONES J. W. (2014). Harmonization and translation of crop modeling data to ensure interoperability. *Environmental Modelling & Software*, **62**, 495–508.

PRESUTTI V., DAGA E., GANGEMI A. & BLOMQVIST E. (2009). extreme design with content ontology design patterns. In *Proc. Workshop on Ontology Patterns*.

RAIMOND Y. & ABDALLAH S. (2007). *The event ontology*. Rapport interne, Citeseer.

RIJGERSBERG H., VAN ASSEM M. & TOP J. (2013). Ontology of units of measure and related concepts. *Semantic Web*, **4**(1), 3–13.

ROUSSEY C., BERNARD S., PINET F., REBOUD X., CELLIER V., SIVADON I., SIMONNEAU D. & BOURIGAULT A.-L. (2017). A methodology for the publication of agricultural alert bulletins as LOD. *Computers and Electronics in Agriculture*, **142**, 632 – 650.

SANDERSON R., CICCARESE P., VAN DE SOMPEL H., BRADSHAW S., BRICKLEY D., CASTRO L. J. G., CLARK T., COLE T., DESENNE P., GERBER A. *et al.* (2013). Open annotation data model. *W3C community draft*, **8**.

SHAW R., TRONCY R. & HARDMAN L. (2009). Lode : Linking open descriptions of events. In *Asian Semantic Web Conference*, p. 153–167 : Springer.

SUÁREZ-FIGUEROA M. C., GÓMEZ-PÉREZ A., MOTTA E. & GANGEMI A. (2012). *Ontology engineering in a networked world*. Springer.

SUÁREZ-FIGUEROA M. C., GÓMEZ-PÉREZ A. & VILLAZÓN-TERRAZAS B. (2009). How to write and use the ontology requirements specification document. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, p. 966–982 : Springer.

UREN V., CIMIANO P., IRIA J., HANDSCHUH S., VARGAS-VERA M., MOTTA E. & CIRAVEGNA F. (2006). Semantic annotation for knowledge management : Requirements and a survey of the state of the art. *Web Semantics : Science, Services and Agents on the World Wide Web*, **4**(1), 14 – 28.

WEIBEL S., KUNZE J., LAGOZE C. & WOLF M. (1998). *Dublin core metadata for resource discovery*. Rapport interne.

# Extraction de connaissances pour aider la décision

Alors que se multiplient des applications tirant profit des avis des utilisateurs et/ou des internautes pour faire des analyses d'usages, de la recommandation, et plus largement assister l'opérateur dans sa prise de décision, le besoin en analyse fine de ces avis s'impose. On peut ainsi classifier certains messages en fonction de différents critères (sujet abordé, niveau d'expertise, niveau informationnel, ou émotions exprimées, par exemple), étudier l'intensité des contradictions exprimées, et en fonction des résultats de ces analyses, conseiller ou aiguiller les utilisateurs. Ces techniques d'analyse peuvent également être mises en œuvre au sein d'applications dédiées, où le retour d'expérience et l'avis d'experts peuvent conjointement assister l'opérateur en situation de crise, par exemple en situation d'urgence humanitaire. À partir de différents contextes applicatifs, les quatre articles de cette session discutent de telles approches, de leurs limites et surtout des perspectives qu'elles ouvrent.

# Prédire l'intensité de contradiction dans les commentaires : faible, forte ou très forte ?

Ismail Badache, Sébastien Fournier, Adrian-Gabriel Chifu

AIX MARSEILLE UNIV, UNIVERSITÉ DE TOULON, CNRS, LIS, MARSEILLE, FRANCE
{ismail.badache, sebastien.fournier, adrian.chifu}@lis-lab.fr

**Résumé** : Les commentaires sur des ressources Web (ex. des cours, des films) deviennent de plus en plus exploitées dans des tâches d'analyse de texte (ex. détection d'opinion, détection de controverses). Cet article étudie l'intensité de contradiction dans les commentaires en exploitant différents critères tels que la variation des notations et la variation des polarités autour d'entités spécifiques (ex. aspects, sujets). Premièrement, les aspects sont identifiés en fonction des distributions des termes émotionnels à proximité des noms les plus fréquents dans la collection des commentaires. Deuxièmement, la polarité est estimée pour chaque segment de commentaire contenant un aspect. Seules les ressources ayant des commentaires contenant des aspects avec des polarités opposées sont prises en compte. Enfin, les critères sont évalués, en utilisant des algorithmes de sélection d'attributs, pour déterminer leur impact sur l'efficacité de la détection de l'intensité des contradictions. Les critères sélectionnés sont ensuite introduits dans des modèles d'apprentissage pour prédire l'intensité de contradiction. L'évaluation expérimentale est menée sur une collection contenant 2244 cours et leurs 73873 commentaires, collectés à partir de *coursera.org*. Les résultats montrent que la variation des notations, la variation des polarités et la quantité de commentaires sont les meilleurs prédicteurs de l'intensité de contradiction. En outre, J48 est l'approche d'apprentissage la plus efficace pour cette tâche.

**Mots-clés** : Analyse de sentiments, Détection d'aspects, Évaluation des critères, Intensité de contradiction.

## 1 Introduction

Au cours des dernières années, le web 2.0 est devenu un espace ouvert où les gens peuvent exprimer leurs opinions en laissant des traces (par exemple, un commentaire, une notation, un j'aime) sur les ressources Web. De nombreux services, tels que les blogs et les réseaux sociaux, représentent une source riche de ces données sociales, qui peuvent être analysées et exploitées dans diverses applications et contextes Badache & Boughanem (2014, 2017a,b). En particulier la détection d'opinion et l'analyse de sentiments Htait *et al.* (2016), par exemple, pour connaître l'attitude d'un client vis-à-vis d'un produit ou de ses caractéristiques, ou pour révéler la réaction des gens à un événement. De tels problèmes nécessitent une analyse rigoureuse des aspects couverts par le sentiment pour produire un résultat représentatif et ciblé.

Une autre problématique concerne la diversité des opinions sur un sujet donné. Certains travaux l'abordent dans le contexte de différents domaines de recherche, avec une notion différente dans chaque cas. Par exemple, Wang & Cardie (2014) visent à identifier des sentiments au niveau d'une phrase exprimée au cours d'une discussion et à les utiliser comme des caractéristiques dans un classifieur qui prédit la dispute dans la discussion. Qiu *et al.* (2013) identifient automatiquement les débats entre des utilisateurs à partir du contenu textuel (interactions) dans les forums, en se basant sur des modèles de variables latentes. Il y a eu d'autres travaux dans l'analyse des interactions avec les utilisateurs, par exemple, l'extraction des expressions de type *agreement* et *disagreement* Mukherjee & Liu (2012) et d'en déduire les relations de l'utilisateur en regardant leurs échanges textuels Awadallah *et al.* (2012).

Cet article étudie les entités (par exemple, les aspects, les sujets) pour lesquelles des contradictions peuvent apparaître dans les commentaires associés à une ressource Web (par exemple des films, des cours) et comment estimer leur intensité. L'intérêt d'estimer l'intensité de la contradiction dépend du cadre d'application. Par exemple, suivre des événements ou des crises politiques controversés tels que la reconnaissance par les États-Unis de Jérusalem comme capitale d'Israël. Cela a généré des opinions (avis) contradictoires, dans les réseaux sociaux, entre différentes communautés à travers le monde. L'estimation de l'intensité de ce conflit peut être utile pour mieux analyser la tendance et les conséquences de cette décision

politique. Dans le cas de la recherche d'information sociale, pour certains besoins d'information, mesurer l'intensité de la contradiction peut être utile pour identifier et classer les documents les plus controversés (par exemple les nouvelles, les événements, etc.). Dans notre cas, connaître l'intensité des opinions contradictoires sur un aspect spécifique (par exemple, *Lecturer, Speaker, Slide, Quiz*) d'un cours en ligne (en anglais) peut être utile pour savoir s'il y a certains éléments à améliorer dans ce cours. Table 1 présente une instance de commentaires contradictoires à propos de l'aspect *Speaker* (conférencier) d'un cours donné.

| Ressource | Commentaire (gauche) | Aspect | Commentaire (droite) | Polarité | Notation |
|---|---|---|---|---|---|
| Cours [1] | The lecturer was an annoying | speaker | and very repetitive. | -0.9 | 1 |
| | Passionate | speaker | and truly amazing things to learn | +0.7 | 4 |

TABLE *1 – Exemple de deux opinions contradictoires sur le "Speaker" d'un cours coursera*

Afin de concevoir notre approche, des tâches fondamentales sont effectuées. Premièrement, l'extraction automatique des aspects caractérisants ces commentaires. Deuxièmement, l'identification des opinions opposées autour de chacun de ces aspects en utilisant un modèle d'analyse des sentiments. Enfin, nous allons évaluer l'impact de certains critères (par exemple, le nombre de commentaires négatifs, le nombre de commentaires positifs) sur l'estimation de l'intensité de contradiction. Plus précisément, nous tentons de sélectionner les critères les plus efficaces et de les combiner avec des approches d'apprentissage pour prédire l'intensité de contradiction. Les principales contributions abordées dans cet article sont doubles :

- **(C1).** Une contradiction dans des commentaires liés à une ressource Web donnée signifie des opinions contradictoires exprimées sur un aspect spécifique, qui est une forme de diversité de sentiments autour de l'aspect au sein de la même ressource. Mais en plus de détecter la contradiction, il est souhaitable d'estimer son intensité. Par conséquent, nous essayons de répondre aux questions de recherche suivantes :
  - ◇ **QR1.** Comment estimer/prédire l'intensité de la contradiction ?
  - ◇ **QR2.** Quel est l'impact de la prise en compte des polarités et des notations sur la prédiction de l'intensité des commentaires contradictoires ?
- **(C2).** La construction d'une collection de test issue du site Web des MOOC [2] `coursera.org`. Cette collection est utile pour l'évaluation des systèmes mesurant l'intensité de contradiction. Des études expérimentales orientées utilisateurs - *user studies* - ont été menées pour collecter les jugements de l'intensité de contradiction (*Not Contradictory, Very Low, Low, Strong* and *Very Strong*).

L'article est organisé comme suit. La section 2 présente certains travaux connexes. La section 3 détaille notre approche pour la prédiction de l'intensité des contradictions autour de certains aspects spécifiques. L'évaluation expérimentale est présentée dans la section 4. Enfin, la section 5 conclut l'article en annonçant des perspectives.

## 2 Vue d'ensemble : État de l'art

La détection de contradictions est un processus complexe qui nécessite souvent l'utilisation de plusieurs méthodes. Plusieurs travaux ont été proposés pour ces méthodes (détection des aspects, analyse de sentiments) mais à notre connaissance, très peu de travaux traitent de la détection et de la mesure de l'intensité de la contradiction. Dans cette section, nous allons brièvement présenter quelques approches de détection de controverses proches de nos travaux puis nous allons présenter les approches liées à la détection des aspects et l'analyse de sentiments, qui sont utiles pour introduire notre approche.

1. `https://www.coursera.org/learn/dog-emotion-and-cognition`
2. Massive Open Online Course

### 2.1 Approches de détection des contradictions et des controverses

Les études les plus liées à notre approche incluent Harabagiu *et al.* (2006), de Marneffe *et al.* (2008), Tsytsarau *et al.* (2010) et Tsytsarau *et al.* (2011), qui tentent de détecter une contradiction dans le texte. Il y a deux approches principales, où les contradictions sont définies comme une forme d'inférence textuelle (par exemple, *entailment identification*) et analysées en utilisant des technologies linguistiques. Harabagiu *et al.* (2006) ont proposé une approche d'analyse des contradictions en exploitant des caractéristiques linguistiques et sémantiques (ex. typologie de verbes), ainsi que des informations syntaxiques telles que la négation (ex. *I love you - I do not love you*) ou l'antonyme (des mots qui ont des significations opposées, c.-à-d. *hot-cold* ou *light-dark*). Leur travail définit les contradictions comme une implication textuelle (*textual entailment* [3]) qui est fausse, lorsque deux phrases expriment des informations mutuellement exclusives sur le même sujet. L'antonymie peut donner lieu à une contradiction lorsque les gens utilisent ces mots pour décrire un sujet.

Poursuivant l'amélioration des travaux dans ce sens, de Marneffe *et al.* (2008) a introduit une classification des contradictions consistant en 7 types qui se distinguent par les caractéristiques qui contribuent à une contradiction, par exemple, l'antonyme, la négation, les discordances numériques qui peuvent être causées par des données erronées : «*there are 7 wonders of the world - the number of wonders of the world are 9*». Ils ont défini les contradictions comme une situation où il est extrêmement improbable que deux phrases soient vraies lorsqu'elles sont ensemble. Tsytsarau *et al.* (2010), (2011) ont proposé une solution automatique et évolutive pour le problème de détection de contradictions en utilisant l'analyse des sentiments. L'intuition de leur approche est que lorsque la valeur agrégée des sentiments (sur un sujet et un intervalle de temps spécifiques) est proche de zéro, alors que la diversité des sentiments est élevée, la contradiction devrait être élevée.

Un autre thème lié à notre travail concerne la détection des controverses et des disputes. Dans la littérature, la détection des controverses a été abordée à la fois par des méthodes supervisées comme dans Popescu & Pennacchiotti (2010), Balasubramanyan *et al.* (2012) et Wang *et al.* (2014) ou par des méthodes non supervisées comme dans Badache *et al.* (2017), Dori-Hacohen & Allan (2015), Garimella *et al.* (2016) et Jang *et al.* (2016). Pour détecter les événements controversés sur Twitter (par exemple, l'accusation de viol de David Copperfield entre 2007 et 2010) [4], Popescu & Pennacchiotti (2010) ont proposé un classifieur basé sur un apprentissage par arbre de décision et un ensemble de caractéristiques telles que les parties du discours, la présence de mots issus du lexique d'opinion ou de controverse, et les interactions des utilisateurs (*retweet* et *reply*). Balasubramanyan *et al.* (2012) ont étendu le modèle LDA (Latent Dirichlet Allocation) supervisé pour prédire comment les membres des différentes communautés politiques réagiront émotionnellement au même sujet c-à-d. la prédiction du niveau de controverse associé à ce sujet. Des classifieurs de type machine à vecteurs de support et régression logistique ont également été proposés par Wang *et al.* (2014) et par Wang & Cardie (2014) pour détecter les disputes dans les discussions sur la page de Wikipedia. Par exemple dans le cas des commentaires sur les modifications des pages Wikipedia [5].

D'autres travaux ont également exploité Wikipédia pour détecter et identifier des sujets controversés sur le Web Dori-Hacohen & Allan (2015), Jang & Allan (2016) et Jang *et al.* (2016). Dori-Hacohen & Allan (2015) et Jang & Allan (2016) ont proposé d'aligner les pages Web aux pages de Wikipedia en supposant qu'une page traite un sujet controversé si la page Wikipedia décrit un sujet lui-même controversé. La nature controversée ou non controversée d'une page Wikipedia est automatiquement détectée sur la base des métadonnées et des discussions associées à la page. Jang *et al.* (2016) ont construit un modèle de langage des sujets controversés appris sur des articles de Wikipédia et utilisé ensuite pour identifier si une page Web est controversée.

La détection des controverses dans les médias sociaux a également été abordée sans supervision en se basant sur les interactions entre les différents utilisateurs Garimella *et al.*

3. https://en.wikipedia.org/wiki/Textual_entailment
4. http://news.bbc.co.uk/2/hi/entertainment/8456070.stm
5. https://www.wikipedia.org/

(2016). Garimella *et al.* (2016) ont proposé d'autres approches de mesure de contradiction basées sur la topologie du réseau, telles que la marche aléatoire (*random walk*), la centralité intermédiaire (*betweenness centrality*) et le plongement de graphe à faible dimension (*low-dimensional graph embeddings*. Les auteurs ont testé des méthodes simples basées sur le contenu et ont noté leur inefficacité par rapport aux méthodes basées sur un graphe utilisateur. D'autres études tentent de détecter des controverses sur des domaines spécifiques, par exemple dans les news Tsytsarau *et al.* (2014) ou dans l'analyse du débat Qiu *et al.* (2013).

Cependant, à notre connaissance, aucun travail antérieur n'a abordé, de manière explicite et concrète, l'intensité de la contradiction ou de la controverse. Dans cet article, contrairement aux travaux antérieurs, plutôt que d'identifier seulement la controverse autour d'un sujet choisi au préalable (par exemple, aspect lié aux nouvelles politiques), nous nous concentrons également sur l'estimation de l'intensité des opinions contradictoires autour de sujets spécifiques. Nous proposons de mesurer l'intensité de la contradiction en utilisant certaines caractéristiques (par exemple, la notation et la polarité).

### 2.2 Approches de détection des aspects

Les premières tentatives de détection d'aspects ont été basées sur l'approche classique d'extraction d'information (IE) en exploitant les phrase nominales fréquentes Hu & Liu (2004). De telles approches fonctionnent bien dans la détection des aspects qui sont sous la forme d'un seul nom, mais sont moins efficaces lorsque les aspects sont de faible fréquence. Dans le contexte de la détection d'aspects, bon nombre de travaux utilisent les CRF (*Conditional Random Fields*) ou les HMM (*Hidden Markov Models*). Parmi ces travaux, nous pouvons citer Hamdan *et al.* (2015) qui utilisent les CRF. D'autres méthodes sont non supervisées et ont prouvé leur efficacité tel que Titov & McDonald (2008) qui construisent un modèle thématique à grains multiples (Multi-Grain Topic Model). Nous pouvons aussi citer le modèle HASM (*unsupervised Hierarchical Aspect Sentiment Model*) proposé par Kim *et al.* (2013) qui permet de découvrir une structure hiérarchique du sentiment fondée sur les aspects dans les avis en ligne non labellés. Dans nos travaux, nous nous sommes inspirés de la méthode non supervisée développée par Poria *et al.* (2014) basée sur l'utilisation de règles d'extraction pour les avis sur les produits. Cette méthode est en cohérence avec nos données expérimentales issues de *coursera.org*.

### 2.3 Approches d'analyse de sentiments

L'analyse du sentiment a fait l'objet de très nombreuses recherches antérieures. Comme dans le cas de la détection d'aspects, les approches supervisées et non supervisées ont chacune leurs solutions. Ainsi, dans les approches non supervisées, nous pouvons citer les approches basées sur les lexiques telles que l'approche développée par Turney (2002) ou bien des méthodes basées sur des corpus comme les travaux de Mohammad *et al.* (2013). Au rang des approches supervisées, nous pouvons citer Pang *et al.* (2002) qui comme nombre de travaux perçoivent la tâche d'analyse de sentiments comme une tâche de classification et utilisent donc des méthodes comme les SVM (*Support Vector Machines*) ou les réseaux bayésiens. D'autres travaux récents sont basés sur les RNN (*Recursive Neural Network*) tels que les travaux de Socher *et al.* (2013). Comme le propos de cet article est de mesurer l'intensité de contradiction et que l'analyse de sentiments n'est qu'une étape du processus, nous avons utilisé l'approche proposée par Radford *et al.* (2017), dont son implémentation est publiquement disponible [6]. Nous décrivons cette méthode dans la section 3.1.2.

## 3 Notre approche : Prédiction de l'intensité des contradictions

Notre approche est basée à la fois sur la détection d'aspects dans les commentaires ainsi que sur l'analyse des sentiments du texte autour de ces aspects. En plus de la détection de

6. `https://github.com/openai/generating-reviews-discovering-sentiment`

contradiction, notre objectif est de prédire le niveau d'intensité de la contradiction en utilisant certains critères et caractéristiques. Ces caractéristiques sont liées à la notation et à la polarité des *commentaires-aspect* (texte autour d'un aspect donné).

### 3.1 Pré-traitement : Identification des polarités autour des aspects

Le pré-traitement est une étape clé pour l'analyse des commentaires (aspects et sentiments). Le module de pré-traitement se compose de trois étapes principales : d'une part, le marquage des termes (identification des noms, verbes, etc), par une analyse syntaxique, au sein des commentaires. Deuxièmement, les noms les plus fréquents dans l'ensemble des commentaires des différents documents sont extraits. Troisièmement, uniquement les noms entourés par des termes émotionnels sont considérés comme des aspects. Nous détaillons ces étapes dans ce qui suit.

#### 3.1.1 Extraction des aspects

Dans notre étude, un aspect est une entité nominale fréquente dans les commentaires et entourée par des termes émotionnels. Afin d'extraire les aspects à partir du texte des commentaires, nous nous sommes basés sur le travail de Poria *et al.* (2014). Cette méthode correspond à nos données expérimentales (commentaires issus de *coursera*). De plus, les traitements suivants sont appliqués :

1. Calcul fréquentiel des termes constituant le corpus des commentaires,
2. Catégorisation des termes (Part-of-speech tagging) de chaque commentaire en utilisant *Stanford Parser* [7],
3. Sélection des termes ayant la catégorie nominale (NN, NNS) [8],
4. Sélection des noms avec des termes émotionnels dans leur voisinage de 5 mots (en utilisant *SentiWordNet* [9]). Le choix de 5 mots a été fait après plusieurs expérimentations,
5. Extraction des termes les plus fréquents (utilisés) dans le corpus parmi ceux sélectionnés dans l'étape précédente. Ces termes seront considérés comme des aspects.

*Exemple* : Soit $C = \{c_1, c_2, c_3\}$ un ensemble de 3 commentaires associés à un document $D$. Nous voulons extraire les aspects à partir de chacun des commentaires en appliquant les étapes décrites ci-dessus.
Nous avons $c_1$ =*"The lecturer was an annoying speaker and very repetitive. I just couldn't listen to him. . . I'm sorry. There was also so much about human development etc that I started to wonder when the info about dogs would start. . . . I found the formatting so different from other courses I've taken, that it was hard to get started and figure things out. Adding to that, was the constant interruption of the "paid certificate" page. If I answer "no" once, please leave me alone ! I also think it's a bit suspect for a prof to be plugging his own book for one of these courses."*

La table 2 récapitule les 5 étapes. Premièrement, nous calculons les fréquences des termes dans l'ensemble des commentaires (à titre d'exemple, les termes "course", "material", "assignments", "content", "lecturer" apparaissent 44219, 3286, 3118, 2947, 2705, respectivement). Deuxièmement, nous étiquetons grammaticalement chaque mots (par exemple, "NN", "NNS" signifient nom en singulier et nom en pluriel, respectivement [10]). Troisièmement, seul les termes de catégorie nominale sont sélectionnés. Quatrièmement, nous gardons uniquement les noms entourés par des termes appartenant au dictionnaire *SentiWordNet* (The *lecturer* was an annoying *speaker* and very repetitive). Enfin, nous considérons comme aspects utiles uniquement les noms qui figurent parmi les noms les plus fréquents dans le corpus des commentaires (l'aspect utile dans ce commentaire est *lecturer*).

7. `http://nlp.stanford.edu:8080/parser/`
8. `https://cs.nyu.edu/grishman/jet/guide/PennPOS.html`
9. `http://sentiwordnet.isti.cnr.it/`
10. `http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html`

| Étape | Description |
|---|---|
| *(1)* | course : 44219, material : 3286, assignments : 3118, content : 2947, lecturer : 2705, ..... terme$_i$ |
| *(2)* | The/DT **lecturer**/NN was/VBD an/DT annoying/VBG **speaker**/NN and/CC very/RB repetitive/JJ ./. I/PRP just/RB could/MD n't/RB listen/VB to/TO him/PRP .../ : I/PRP 'm/VBP sorry/JJ ./. There/EX was/VBD also/RB so/RB much/JJ about/IN human/JJ **development**/NN **etc**/NN that/IN I/PRP started/VBD to/TO wonder/VB when/WRB the/DT **info**/NN about/IN **dogs**/NNS would/MD start/VB .../ : ./. I/PRP found/VBD the/DT **formatting**/NN so/RB different/JJ from/IN other/JJ **courses**/NNS I/PRP 've/VBP taken/VBN ,/, that/IN it/PRP was/VBD hard/JJ to/TO get/VB started/VBN and/CC figure/VB **things**/NNS out/RP ./. Adding/VBG to/TO that/DT ,/, was/VBD the/DT constant/JJ **interruption**/NN of/IN the/DT "/" paid/VBN **certificate**/NN "/" **page**/NN ./. If/IN I/PRP answer/VBZ "/" no/UH "/" once/RB ,/, please/VB leave/VB me/PRP alone/RB !/. I/PRP also/RB think/VBP it/PRP 's/VBZ a/DT bit/RB suspect/JJ for/IN a/DT **prof**/NN to/TO be/VB plugging/VBG his/PRP$ own/JJ **book**/NN for/IN one/CD of/IN these/DT **courses**/NNS ./. |
| *(3)* | lecturer, speaker, development, dogs, formatting, courses, interruption, certificate, page, prof |
| *(4)* | lecturer, speaker |
| *(5)* | lecturer |

TABLE *2 – Les différentes étapes pour extraire les aspects dans un commentaire*

Une fois que nous avons défini la liste des aspects qui caractérisent notre collection de données, nous devons estimer la polarité des sentiments autour de ces aspects. La section suivante présente notre modèle d'analyse de sentiments.

#### 3.1.2 Analyse de sentiments

Le sentiment porté par un commentaire sur un aspect donné (commentaire-aspect) est estimé en utilisant la méthode appelée *SentiNeuron* [11]. *SentiNeuron* est un modèle non supervisé proposé par Radford *et al.* (2017) pour détecter les signaux de sentiment dans les commentaires. Cette approche est basée sur les réseaux de neurones récurrent (*recurrent neural network*) de type mLSTM (multiplicative Long Short-Term Memory). Radford *et al.* (2017) ont également trouvé qu'une unité dans le mLSTM correspond directement au sentiment de la sortie. Les auteurs ont mené une série d'expérimentations sur plusieurs collections de tests telles que les collections des commentaires issues d'Amazon McAuley *et al.* (2015) et d'IMDb [12]. Cette approche fournie une précision de 91.8%, et surpasse de manière significative plusieurs approches de l'état de l'art telles que celles présentées dans Looks *et al.* (2017). Nous notons que le terme polarité signifie sentiment, c'est une valeur comprise entre -1 et 1.

### 3.2 Collection de test : coursera.org

#### 3.2.1 Données collectées

A notre connaissance, il n'existe pas à ce jour de collection de test standard, contenant des informations comme les aspects, les notations et les polarités des commentaires, pour évaluer l'efficacité des systèmes de détection de contradictions dans les commentaires. De ce fait, dans le but d'expérimenter l'efficacité de notre approche, nous avons collecté 2244 ressources en anglais extraites du site "coursera.org" via son API [13], entre le 10 et le 14 octobre 2016.

11. https://github.com/openai/generating-reviews-discovering-sentiment
12. https://www.cs.cornell.edu/personnes/pabo/film-review-data/
13. https://building.coursera.org/app-platform/catalog

Chaque ressource décrit un cours et est représentée par un ensemble de métadonnées. Pour chaque cours, nous avons collecté également ses commentaires et ses notations via le *parsing* des pages web des cours (voir les statistiques sur la table 3).

| Champ | Nombre |
|---|---|
| Cours | 2244 |
| Cours notés | 1115 |
| Commentaires | 73873 |
| notations | 298326 |
| Commentaires avec notation ★☆☆☆☆ | 1705 |
| Commentaires avec notation ★★☆☆☆ | 1443 |
| Commentaires avec notation ★★★☆☆ | 3302 |
| Commentaires avec notation ★★★★☆ | 12202 |
| Commentaires avec notation ★★★★★ | 55221 |

TABLE *3 – Les chiffres des données de la collection de Coursera.org*

Nous avons pu capturer automatiquement 22 aspects utiles à partir de l'ensemble des commentaires (voir table 4). La table 4 présente les statistiques sur les 22 aspects détectés, par exemple, pour l'aspect *Slide* nous avons enregistré : 56 notations d'une étoile, 64 notations de deux étoiles, 81 notations de trois étoiles, 121 notations de quatre étoiles, 115 notations de cinq étoiles, 131 commentaires avec une polarité négative, 102 commentaires avec une polarité positive ainsi que 192 commentaires et 41 cours concernant cet aspect.

| Aspects | Not1 | Not2 | Not3 | Not4 | Not5 | Négatif | Positif | Comment | Cours |
|---|---|---|---|---|---|---|---|---|---|
| Assignment | 204 | 208 | 333 | 840 | 1726 | 1057 | 1763 | 2384 | 186 |
| Content | 176 | 179 | 341 | 676 | 1641 | 505 | 1496 | 1883 | 207 |
| Exercise | 29 | 46 | 94 | 290 | 693 | 195 | 531 | 673 | 58 |
| Information | 100 | 123 | 238 | 523 | 1389 | 299 | 1165 | 1359 | 143 |
| Instructor | 129 | 106 | 122 | 302 | 1514 | 295 | 1107 | 1322 | 140 |
| Knowledge | 74 | 72 | 121 | 400 | 1604 | 905 | 791 | 1243 | 178 |
| Lecture | 185 | 206 | 290 | 613 | 1762 | 763 | 1508 | 1988 | 208 |
| Lecturer | 32 | 41 | 48 | 85 | 461 | 55 | 193 | 236 | 39 |
| Lesson | 40 | 59 | 75 | 224 | 712 | 187 | 420 | 554 | 84 |
| Material | 191 | 203 | 328 | 722 | 2234 | 784 | 1693 | 2254 | 237 |
| Method | 19 | 23 | 40 | 125 | 404 | 53 | 187 | 224 | 31 |
| Presentation | 46 | 50 | 75 | 142 | 413 | 93 | 196 | 274 | 54 |
| Professor | 76 | 74 | 129 | 452 | 3001 | 331 | 2234 | 2369 | 151 |
| Quality | 55 | 53 | 51 | 110 | 372 | 113 | 170 | 262 | 54 |
| Question | 94 | 98 | 172 | 284 | 356 | 311 | 289 | 502 | 104 |
| Quiz | 151 | 155 | 221 | 401 | 581 | 481 | 475 | 824 | 128 |
| Slide | 56 | 64 | 81 | 121 | 115 | 131 | 102 | 192 | 47 |
| Speaker | 17 | 15 | 34 | 70 | 170 | 34 | 72 | 103 | 24 |
| Student | 140 | 105 | 171 | 383 | 1035 | 519 | 709 | 1066 | 172 |
| Teacher | 62 | 46 | 82 | 293 | 2180 | 248 | 1481 | 1642 | 119 |
| Topic | 67 | 89 | 176 | 437 | 1154 | 236 | 951 | 1066 | 130 |
| Video | 228 | 238 | 356 | 707 | 1614 | 941 | 1421 | 2058 | 245 |
| Nombre total : 22 aspects détectés | | | | | | | | | |

TABLE *4 – Statistiques sur les aspects issus des commentaires de Coursera.org*

### 3.2.2 Jugements par les utilisateurs (contradictions et sentiments)

Afin d'obtenir des jugements de contradictions et de sentiments pour un aspect donné : 1) nous avons demandé à trois utilisateurs d'évaluer la classe de sentiment pour chaque

commentaires-aspect de 1100 cours ; 2) trois autres utilisateurs ont évalué le degré de contradiction entre les commentaires-aspect. En moyenne 60 commentaires-aspect par cours sont jugés manuellement pour chaque aspect (totalement : 66104 commentaires-aspect de 1100 cours, c'est-à-dire 50 cours pour chaque aspect). Nous notons que chaque aspect a été jugé par 3 utilisateurs.


FIGURE *1 – Interface du système d'évaluation*

Pour évaluer les sentiments et les contradictions dans les commentaires-aspect de chaque cours, nous utilisons une échelle de notation de 3 points pour les sentiments : (*Negative, Neutral, Positive*) ; et une échelle de 5 points pour les contradictions : *Not Contradictory, Very Low, Low, Strong* et *Very Strong* (voir la figure 1).

Nous avons analysé le degré d'accord entre les évaluateurs des contradictions pour chaque aspect avec la mesure Kappa Cohen $k$ Cohen (1960). Cet indicateur prend en compte la proportion d'accord entre les évaluateurs et la proportion de l'accord attendu entre les évaluateurs par hasard. La mesure de Kappa est égale à $1$ si les évaluateurs sont complètement d'accord, $0$ s'ils ne sont d'accord que par hasard. $k$ est négatif si l'accord entre évaluateurs est pire que l'aléatoire. Comme nous avons trois évaluateurs par aspect, la valeur Kappa a été calculée pour chaque paire d'évaluateurs, puis leur moyenne a été calculée.

La figure 2 montre la distribution de la mesure kappa pour chaque aspect. Nous constatons que la mesure de l'accord varie de $0.60$ à $0.91$. La mesure moyenne d'accord entre les évaluateurs est de 80%, ce qui correspond à un accord fort. Concernant l'analyse du degré d'accord entre les évaluateurs des sentiments, nous avons trouvé un accord de Kappa $k = 0.78$, qui correspond aussi à un accord fort.

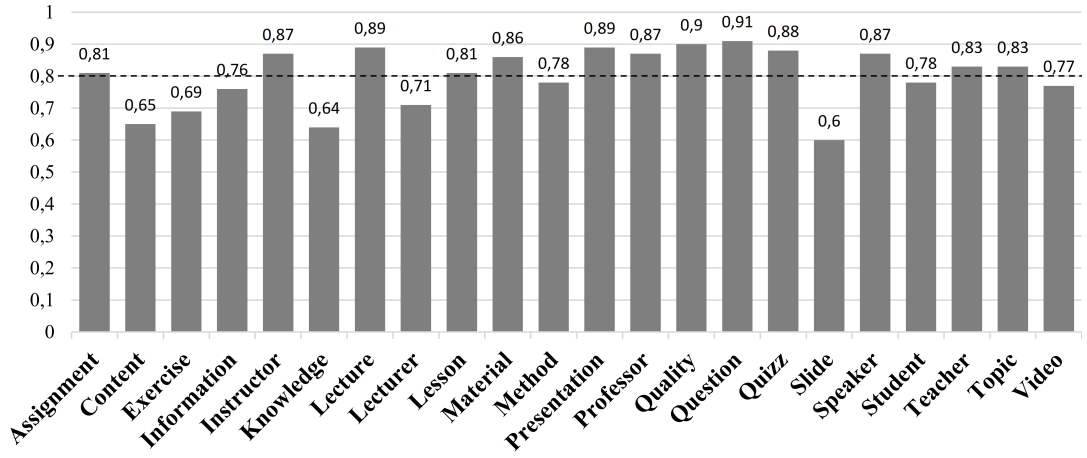



FIGURE *2 – Répartition de la mesure Kappa $k$ par aspect. $< 0$ désaccord, $0.0 - 0.2$ accord très faible, $0.21 - 0.40$ accord faible, $0,41 - 0,6$ accord modéré, $0,61 - 0,80$ accord fort, $0,81 - 1$ accord parfait.*

### 3.3 Identification des critères les plus efficaces

Dans cette étude, nous nous sommes appuyés sur des algorithmes de sélection d'attributs pour déterminer les critères les plus importants pour la tâche de prédiction d'intensité de contradiction. Les algorithmes de sélection d'attributs visent à identifier et supprimer le maximum d'information inutile, redondante et non pertinente en amont d'un processus à base d'apprentissage Hall & Holmes (2003). Ils permettent également de sélectionner de manière automatique les sous ensembles de critères permettant d'avoir les meilleurs résultats. Nous avons utilisé Weka [14] (dernière version stable 2018 : 3.8.2), un outil open-source écrit entièrement en Java et qui rassemble un bon ensemble de techniques d'apprentissage et des techniques de sélection d'attributs.

| $c_i$ | Critère | Description |
|---|---|---|
| $c_1$ | NegCom | Nombre de commentaires négatifs sur le document |
| $c_2$ | PosCom | Nombre de commentaires positifs sur le document |
| $c_3$ | TotalCom | Nombre total des commentaires sur le document |
| $c_4$ | Not1 | Nombre de commentaires avec notation ★☆☆☆☆ |
| $c_5$ | Not2 | Nombre de commentaires avec notation ★★☆☆☆ |
| $c_6$ | Not3 | Nombre de commentaires avec notation ★★★☆☆ |
| $c_7$ | Not4 | Nombre de commentaires avec notation ★★★★☆ |
| $c_8$ | Not5 | Nombre de commentaires avec notation ★★★★★ |
| $c_9$ | VarNot | Variation des notations (écart type selon Pearson & Stephens (1964)) |
| $c_{10}$ | VarPol | Variation des polarités (écart type selon Pearson & Stephens (1964)) |

TABLE *5 – Liste des critères exploités*

La table 5 présente les 10 critères que nous avons considérés pour prédire l'intensité de contradiction dans les commentaires. La nature des critères $c_1$ jusqu'au critère $c_8$ est un simple comptage, par exemple les critères $c_1$ et $c_2$ liés à la polarité représentent le nombre de commentaires négatifs et positifs sur le document, respectivement. Les critères $c_4$, $c_5$, $c_6$, $c_7$ et $c_8$ sont liés à la notation. La notation est une note sur une échelle de 1 à une valeur max de 5, où 3 signifie "moyen" et 5 signifie "excellent". Concernant les deux derniers critères $c_9$ et $c_{10}$, ils représentent la variation des notations et des polarités des commentaires pour un aspect donné associés à un document (un cours dans notre cas). Ces deux critères sont calculé en se basant sur la formule de l'écart type suivante, proposée par Pearson & Stephens (1964) :

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x - \bar{x})^2}{n}} \tag{1}$$

Où $x$ est la valeur du critère (notation, polarité), $\bar{x}$ est la moyenne de l'échantillon du critère concerné, et $n$ est la taille de l'échantillon.

Dans cette étude, nous avons procédé ainsi : 50 cours avec leurs commentaires pour chaque aspect (22 aspects) de la collection *coursera* ont été extraits aléatoirement. Ensuite, nous avons considéré l'échelle des 4 points comme des classes reflétant l'intensité des contradictions autour d'un aspect spécifique : *Very Low, Low, Strong* et *Very Strong*, selon les jugements des évaluateurs. L'ensemble résultant contient 1100 cours (instances) répartis selon leur classe d'intensité de contradiction comme suit :

— 230 Very Low
— 264 Low
— 330 Strong
— 276 Very Strong

14. `http://www.cs.waikato.ac.nz/ml`

Les classes de ces ensembles sont déséquilibrées, or lorsque le nombre d'éléments d'une classe dans une collection d'apprentissage dépasse considérablement les autres échantillons des autres classes, un classifieur tend à prédire les échantillons de la classe majoritaire et peut ignorer complètement les classes minoritaires Yen & Lee (2006). Pour cette raison, nous avons appliqué une approche de sous-échantillonnage (en réduisant le nombre d'échantillons qui ont la classe majoritaire) pour générer des collections équilibrées composées de :

— 230 Very Low

— 230 Low

— 230 Strong

— 230 Very Strong

Les classes *Low, Strong* et *Very Strong* ont été sélectionnées aléatoirement. Enfin, nous avons appliqué les algorithmes de sélection d'attributs sur les quatre ensembles obtenus, pour 5 itérations de validation croisée (5-folds cross-validation).

Dans notre cas, les algorithmes de sélection d'attributs consistent à attribuer un score à chaque critère en fonction de sa signification vis-à-vis la classe d'intensité de contradiction (*Very Low, Low, Strong* et *Very Strong*). Ces algorithmes fonctionnent différemment : certains retournent un classement d'importance des critères (par exemple, *FilteredAttributeEval*), tandis que d'autres retournent le nombre de fois qu'un critère donné a été sélectionné par un algorithme dans une validation croisée (par exemple, *FilteredSubsetEval*). Nous notons que nous avons utilisé pour chaque algorithme le paramétrage par défaut fourni par Weka.

Nous avons appliqué une validation croisée à 5 itérations pour $10$ critères, c'est-à-dire $n = 10$. La table 6 présente les critères sélectionnés par les algorithmes de sélection d'attributs. Nous avons utilisé deux types de ces algorithmes : a) ceux qui utilisent des méthodes de classement pour ordonner les critères sélectionnés (la métrique dans la table est [Rank]) ; et b) ceux qui utilisent des méthodes de recherche qui indiquent combien de fois le critère a été sélectionné pendant la tâche de la validation croisée (la métrique dans la table est [Folds]). Un critère fortement préféré (choisi) par l'algorithme de sélection est un critère bien classé, c'est-à-dire *Rank* $= 1$ et fortement sélectionné, c'est-à-dire *Folds* $= 5$.

| Algorithm | Metric | $c_1$ | $c_2$ | $c_3$ | $c_4$ | $c_5$ | $c_6$ | $c_7$ | $c_8$ | $c_9$ | $c_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CfsSubsetEval | [Folds] | 5 | 5 | 2 | 0 | 0 | 0 | 0 | 0 | 5 | 5 |
| WrapperSubsetEval | [Folds] | 4 | 4 | 4 | 2 | 0 | 0 | 0 | 2 | 5 | 5 |
| ConsistencySubsetEval | [Folds] | 5 | 5 | 4 | 2 | 1 | 1 | 2 | 2 | 5 | 5 |
| FilteredSubsetEval | [Folds] | 5 | 5 | 4 | 3 | 2 | 2 | 3 | 3 | 5 | 5 |
| | Moyenne | 4.75 | 4.75 | 3.5 | 1.75 | 0.75 | 0.75 | 1.25 | 1.75 | 5 | 5 |
| ChiSquaredAttributeEval | [Rank] | 3 | 4 | 5 | 7 | 9 | 10 | 8 | 6 | 2 | 1 |
| FilteredAttributeEval | [Rank] | 4 | 3 | 5 | 7 | 9 | 10 | 8 | 6 | 2 | 1 |
| GainRatioAttributeEval | [Rank] | 3 | 4 | 5 | 7 | 9 | 10 | 8 | 6 | 2 | 1 |
| InfoGainAttributeEval | [Rank] | 3 | 4 | 5 | 7 | 9 | 10 | 8 | 6 | 1 | 2 |
| OneRAttributeEval | [Rank] | 4 | 3 | 5 | 7 | 9 | 10 | 8 | 6 | 2 | 1 |
| ReliefFAttributeEval | [Rank] | 4 | 3 | 6 | 8 | 9 | 10 | 7 | 5 | 1 | 2 |
| SVMAttributeEval | [Rank] | 4 | 3 | 5 | 7 | 9 | 10 | 8 | 6 | 2 | 1 |
| SymetricalUncertEval | [Rank] | 3 | 4 | 5 | 7 | 9 | 10 | 8 | 6 | 2 | 1 |
| | Moyenne | 3.5 | 3.5 | 5.12 | 7.12 | 9 | 10 | 7.87 | 5.87 | 1.75 | 1.25 |

TABLE *6 – Les critères sélectionnés par les algorithmes de sélection d'attributs*

La table 6 montre que les critères $c_{10}$ : *VarPol*, $c_9$ : *VarNot*, $c_1$ : *NegCom* et $c_2$ : *PosCom* sont les plus sélectionnés et les mieux classés par rapport aux autres critères. Les critères $c_3$ : *TotalCom*, $c_4$ : *Not1* et $c_8$ : *Not5* sont modérément favorisés par les algorithmes de sélection d'attributs, à l'exception de l'algorithme *CfsSubsetEval* qui n'a pas sélectionné $c_4$ et $c_8$. Les critères $c_5$, $c_6$ et $c_7$ ne sont pas sélectionnés à la fois par les algorithmes *CfsSubsetEval* et *WrapperSubsetEval*. Enfin, les critères les plus faibles et les plus désavantagés sont $c_5$ : *Not2* et $c_6$ : *Not3*, ils sont ordonnés au rang 9 et 10, respectivement.

### 3.4 Apprentissage des critères pour prédire l'intensité de contradiction

D'autres expérimentations ont été menées en exploitant ces critères dans des approches supervisées basées sur des modèles d'apprentissage. Nous avons utilisé les instances (les cours) des 22 aspects de la collection *coursera.org* comme ensembles d'apprentissage. Nous avons ensuite utilisé trois algorithmes d'apprentissage. Ce choix s'explique par le fait qu'ils ont souvent montré leur efficacité dans les tâches d'analyse de texte : SVM Vosecky *et al.* (2012), J48 (implémentation C4.5) Quinlan (1993) et Naive Bayes Yuan *et al.* (2012). L'entrée de chaque algorithme est un vecteur de critères (voir table 5), soit tous les critères ou seulement les critères sélectionnés par un algorithme de sélection précis. Les algorithmes d'apprentissage prédisent la classe d'intensité de contradiction pour les cours (*Very Low, Low, Strong* et *Very Strong*). Enfin, nous avons appliqué une validation croisée pour 5 itérations (5-folds cross-validation). La figure 3 illustre le processus d'apprentissage que nous avons mis en place pour l'évaluation des critères.

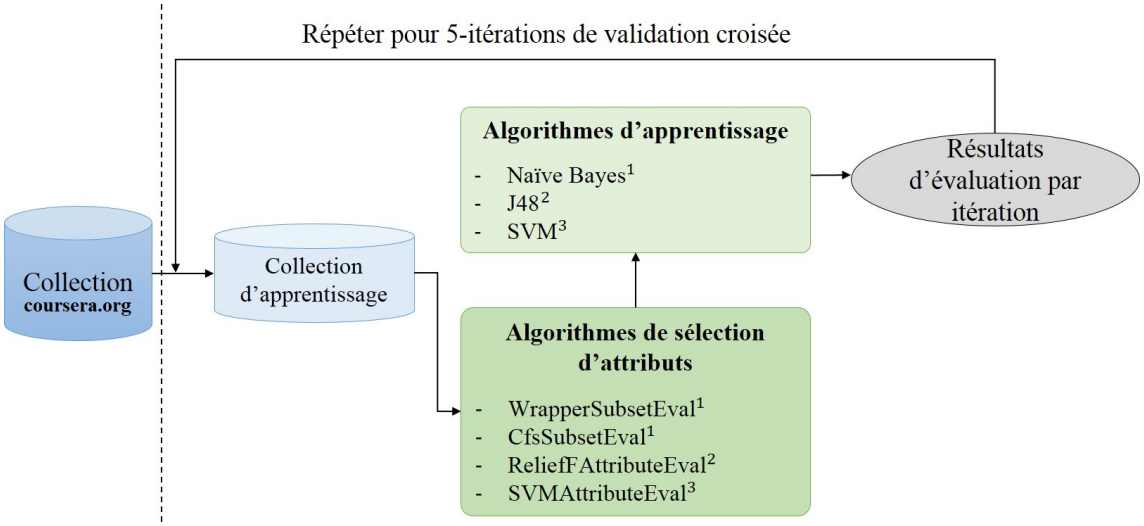



FIGURE *3 – Processus d'apprentissage en utilisant les algorithmes de sélection*

Nous rappelons que la phase des algorithmes de sélection d'attributs a fait ressortir les ensembles de critères suivants (voir la table 7).

| Algorithmes de sélection | Critères |
|---|---|
| CfsSubsetEval | $c_1$, $c_2$, $c_3$, $c_9$, $c_{10}$ |
| WrapperSubsetEval | $c_1$, $c_2$, $c_3$, $c_4$, $c_8$, $c_9$, $c_{10}$ |
| Other algorithms | $c_1$, $c_2$, $c_3$, $c_4$, $c_5$, $c_6$, $c_7$, $c_8$, $c_9$, $c_{10}$ |

TABLE *7 – Ensembles des critères sélectionnés*

La question à ce stade est liée à la spécification du vecteur de critères d'entrée pour les algorithmes d'apprentissage, soit on prend tous les critères, soit on garde uniquement ceux sélectionnés par les techniques de sélection d'attributs. Dans ce cas, avec quels algorithmes d'apprentissage ces derniers seront combinés.

Afin de prendre en compte les critères choisis par les algorithmes de sélection dans des modèles d'apprentissage, nous nous sommes basés sur les travaux de Hall & Holmes (2003). Ils ont étudié l'efficacité de certaines techniques de sélection d'attributs en les confrontant avec les techniques d'apprentissage. Étant donné que la performance des critères diffère d'une technique d'apprentissage à une autre, ils ont identifié les meilleures techniques de sélection d'attributs permettant de retrouver les critères les plus performants en fonction des techniques d'apprentissage à utiliser. En se basant sur leur étude, nous avons utilisé les mêmes couples des techniques d'apprentissage et des techniques de sélection d'attributs :

— L'ensemble des critères sélectionnés par *CfsSubsetEval* (CFS) et *WrapperSubsetEval* (WRP) sont appris par le modèle Naïve Bayes.
— L'ensemble des critères sélectionnés par *ReliefFAttributeEval* (RLF) sont appris par le modèle J48 (C4.5 implementation).
— L'ensemble des critères sélectionnés par *SVMAttributeEval* (SVM) sont appris par le modèle SVM à multi-classes (appelé *SMO function* sur Weka).

Afin de vérifier la significativité des résultats par rapport aux résultats de Naïve Bayes (considérés comme références - résultats de base), nous avons effectué le test de Student Student (1908). Nous avons attaché * (forte signification) et ** (très forte signification) aux résultats de la table 8 quand *p-value <0.05* et *p-value <0.01*, respectivement.

| Classifieurs | Classes (Niveaux d'intensité) | Techniques de sélection | Tous les critères |
|---|---|---|---|
| Naïve Bayes (Baseline) | Very Low | 0.81 (CFS) | 0.71 |
| | Low | 0.38 (CFS) | 0.34 |
| | Strong | 0.75 (CFS) | 0.66 |
| | Very Strong | 0.78 (CFS) | 0.69 |
| | Moyenne | 0.68 (CFS) | 0.60 |
| | Very Low | 0.86 (WRP) | 0.72 |
| | Low | 0.46 (WRP) | 0.38 |
| | Strong | 0.76 (WRP) | 0.63 |
| | Very Strong | 0.80 (WRP) | 0.67 |
| | Moyenne | 0.72 (WRP) | 0.60 |
| SVM | Very Low | $0.88^{*}$ (SVM) | $0.88^{*}$ |
| | Low | $0.72^{**}$ (SVM) | $0.72^{**}$ |
| | Strong | $0.78^{*}$ (SVM) | $0.78^{*}$ |
| | Very Strong | $0.90^{**}$ (SVM) | $0.90^{**}$ |
| | Moyenne | $0.82^{**}$ (SVM) | $0.82^{**}$ |
| J48 | Very Low | $0.97^{**}$ (RLF) | $0.97^{**}$ |
| | Low | $0.92^{**}$ (RLF) | $0.92^{**}$ |
| | Strong | $0.97^{**}$ (RLF) | $0.97^{**}$ |
| | Very Strong | $0.98^{**}$ (RLF) | $0.98^{**}$ |
| | Moyenne | $0.96^{**}$ (RLF) | $0.96^{**}$ |

TABLE *8 – Les résultats de précision Weka pour les techniques d'apprentissage automatique*

La table 8 présente les résultats des trois algorithmes d'apprentissage des critères ressortis de l'étude utilisant les techniques de sélection d'attributs. Les résultats sont discutés ci-dessous pour chaque algorithme d'apprentissage.

### 3.4.1 Résultats obtenus par Naïve Bayes (Baseline)

Les résultats en termes de précision obtenus en utilisant des algorithmes de sélection CFS et WRP avec NaiveBayes, sont de 0.68 et 0.72, respectivement. Ces résultats dépassent ceux obtenus en utilisant tous les critères (précision : 0.60). En effet, nous avons enregistré des taux d'amélioration moyens de $14\%$ et $20\%$ pour Naïve Bayes en utilisant seulement les critères sélectionnés par CFS (0.68) et WRP (0.72), respectivement, par apport au résultat obtenu en utilisant tous les critères (0.60). Par conséquent, les approches d'apprentissage automatique peuvent donner une meilleure efficacité (précision) quand ils sont combinés avec des approches de sélection d'attributs. Les meilleures précisions sont obtenues pour les classes *Very Strong, Strong* et *Very Low*. Il semble que la classe *Low* est difficile à prédire avec Naïve Bayes, tout en utilisant à la fois les deux algorithmes de sélection CFS (0.38) et WRP (0.46).

### 3.4.2 Résultats obtenus par SVM

Les résultats obtenus par SVM en utilisant l'algorithme de sélection *SVMAttributeEval*, où tous les critères ont été sélectionnés, sont meilleurs par rapport à ceux obtenus par Naïve Bayes. Nous avons enregistré des taux d'amélioration moyens de $21\%$ et $14\%$ pour SVM par rapport à Naïve Bayes en utilisant CFS et WRP, respectivement. Nous avons également remarqué que SVM était capable de prédire la classe *Low* avec une meilleure précision que celle fournie par Naïve Bayes. Même si l'algorithme SVM est un peu coûteux en termes de temps d'exécution par rapport à Naïve Bayes, il reste favoriser pour obtenir des résultats significatifs en termes de précision.

### 3.4.3 Résultats obtenus par J48

Les résultats confirment que l'arbre de décision J48 est le modèle le plus approprié, il prend en compte tous les critères de manière plus efficace que les autres configurations. Les taux d'amélioration moyens par rapport à Naïve Bayes (en utilisant CFS et WRP) et SVM sont 41%, 33% et 17%, respectivement. En outre, les améliorations sont également fortement significatives pour chaque classe par rapport à SVM et Naïve Bayes. La classe *Low*, difficile à prédire avec les configurations précédentes, a été prédite avec une très forte précision de 92%. Comparées à Naïve Bayes (en utilisant CFS et WRP) et SVM, les améliorations enregistrées concernant la classe *Low* sont de 142%, 100% et 28%, respectivement.

Enfin, tous ces résultats expérimentaux montrent clairement que l'approche proposée permet de détecter de manière significative l'intensité de la contradiction dans les commentaires. Nous avons constaté que les résultats obtenus, par les deux algorithmes CFS et WRP, confirment l'hypothèse lancé par Hall et Holmes. C'est en effet les deux seuls cas pour lequel les résultats de précision obtenus avec la sélection d'attributs, soient 0.68 (CFS) et 0.72 (WRP), dépassent l'utilisation de tous les critères, 0.60 en termes de précision. Ces améliorations montrent l'intérêt de combiner les algorithmes de sélection d'attributs avec les modèles d'apprentissage. En outre, le modèle J48 a donné les meilleures améliorations par rapport à toutes les autres configurations. Nous concluons que les ressources (cours) ayant des opinions plus diversifiées (commentaires positifs et négatifs), sont susceptibles d'avoir des contradictions avec différents niveaux d'intensité.

## 4 Conclusion

Cet article propose une approche supervisée exploitant un ensemble de critères permettant de prédire l'intensité de la contradiction, en attirant l'attention sur les aspects dans lesquels les utilisateurs ont des opinions contradictoires. L'intuition derrière l'approche proposée est que les notations et les sentiments associés aux commentaires sur un aspect spécifique peuvent être considérés comme des critères (ex. diversité des sentiments et des notations en fonction de l'écart-type) pour mesurer l'intensité de contradiction. L'évaluation expérimentale menée sur la collection issue de *coursera.org* montre que les critères *NegCom, PosCom, VarNot* et *VarPol* sont les plus fructueux pour prédire l'intensité de la contradiction. De plus, les algorithmes d'apprentissage basés sur les critères les plus pertinents selon les algorithmes de sélection d'attributs sont généralement mieux comparés à ceux obtenus lorsque les algorithmes de sélection d'attributs sont ignorés. L'algorithme J48 apporte les meilleurs résultats par rapport à Naïve Bayes et SVM. Enfin, nous notons que nous sommes conscients que l'évaluation de notre approche est encore limitée. La principale faiblesse de notre approche est sa dépendance à la qualité des modèles de sentiments et d'extraction d'aspect. D'autres expérimentations à plus grande échelle sur d'autres types de collections sont également envisagées. Ceci étant même avec ces éléments simples, les premiers résultats obtenus nous encouragent à investir davantage cette piste.

## Références

AWADALLAH A. H., ABU-JBARA A. & RADEV D. R. (2012). Detecting subgroups in online discussions by modeling positive and negative relations among participants. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012, July 12-14, 2012, Jeju Island, Korea*, p. 59–70.

BADACHE I. & BOUGHANEM M. (2014). Harnessing social signals to enhance a search. In *Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT) - Volume 01*, WI-IAT '14, p. 303–309, Washington, DC, USA.

BADACHE I. & BOUGHANEM M. (2017a). Emotional social signals for search ranking. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, p. 1053–1056, New York, NY, USA : ACM.

BADACHE I. & BOUGHANEM M. (2017b). Fresh and diverse social signals : Any impacts on search ? In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, CHIIR '17, p. 155–164, New York, NY, USA : ACM.

BADACHE I., FOURNIER S. & CHIFU A. (2017). Finding and quantifying temporal-aware contradiction in reviews. In *Information Retrieval Technology - 13th Asia Information Retrieval Societies Conference, AIRS 2017, Jeju Island, South Korea, November 22-24, 2017, Proceedings*, p. 167–180.

BALASUBRAMANYAN R., COHEN W. W., PIERCE D. & REDLAWSK D. P. (2012). Modeling polarizing topics : When do different political communities respond differently to the same news ? In *Proceedings of the Sixth International Conference on Weblogs and Social Media, Dublin, Ireland, June 4-7, 2012*.

COHEN J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, **20**(1), 37–46.

DE MARNEFFE M., RAFFERTY A. N. & MANNING C. D. (2008). Finding contradictions in text. In *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA*, p. 1039–1047.

DORI-HACOHEN S. & ALLAN J. (2015). Automated controversy detection on the web. In *Advances in Information Retrieval - 37th European Conference on IR Research, ECIR 2015, Vienna, Austria, March 29 - April 2, 2015. Proceedings*, p. 423–434.

GARIMELLA K., MORALES G. D. F., GIONIS A. & MATHIOUDAKIS M. (2016). Quantifying controversy in social media. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, San Francisco, CA, USA, February 22-25, 2016*, p. 33–42.

HALL M. A. & HOLMES G. (2003). Benchmarking attribute selection techniques for discrete class data mining. *IEEE Trans. Knowl. Data Eng.*, **15**(6), 1437–1447.

HAMDAN H., BELLOT P. & BÉCHET F. (2015). Lsislif : CRF and logistic regression for opinion target extraction and sentiment polarity analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT, Colorado, USA, June 4-5, 2015*, p. 753–758.

HARABAGIU S. M., HICKL A. & LACATUSU V. F. (2006). Negation, contrast and contradiction in text processing. In *Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, July 16-20, 2006, Boston, Massachusetts, USA*, p. 755–762.

HTAIT A., FOURNIER S. & BELLOT P. (2016). LSIS at semeval-2016 task 7 : Using web search engines for english and arabic unsupervised sentiment intensity prediction. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, p. 469–473.

HU M. & LIU B. (2004). Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004*, p. 168–177.

JANG M. & ALLAN J. (2016). Improving automated controversy detection on the web. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, p. 865–868.

JANG M., FOLEY J., DORI-HACOHEN S. & ALLAN J. (2016). Probabilistic approaches to controversy detection. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*, p. 2069–2072.

KIM S., ZHANG J., CHEN Z., OH A. H. & LIU S. (2013). A hierarchical aspect-sentiment model for online reviews. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, July 14-18, 2013, Bellevue, Washington, USA*.

LOOKS M., HERRESHOFF M., HUTCHINS D. & NORVIG P. (2017). Deep learning with dynamic computation graphs. *CoRR*, **abs/1702.02181**.

MCAULEY J. J., PANDEY R. & LESKOVEC J. (2015). Inferring networks of substitutable and complementary products. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*, p. 785–794.

MOHAMMAD S., KIRITCHENKO S. & ZHU X. (2013). Nrc-canada : Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the 7th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2013, Atlanta, Georgia, USA, June 14-15, 2013*, p. 321–327.

MUKHERJEE A. & LIU B. (2012). Mining contentions from discussions and debates. In *The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, Beijing, China, August 12-16, 2012*, p. 841–849.

PANG B., LEE L. & VAITHYANATHAN S. (2002). Thumbs up ? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, EMNLP 2002, Philadelphia, PA, USA, July 6-7, 2002*.

PEARSON E. & STEPHENS M. (1964). The ratio of range to standard deviation in the same normal sample. *Biometrika*, **51**(3/4), 484–487.

POPESCU A. & PENNACCHIOTTI M. (2010). Detecting controversial events from twitter. In *Proceedings of the 19th ACM Conference on Information and Knowledge Management, CIKM 2010, Toronto, Ontario, Canada, October 26-30, 2010*, p. 1873–1876.

PORIA S., CAMBRIA E., KU L., GUI C. & GELBUKH A. F. (2014). A rule-based approach to aspect extraction from product reviews. In *Proceedings of the Second Workshop on Natural Language Processing for Social Media, SocialNLP@COLING, Dublin, Ireland, August 24, 2014*, p. 28–37.

QIU M., YANG L. & JIANG J. (2013). Modeling interaction features for debate side clustering. In *22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013*, p. 873–878.

QUINLAN J. R. (1993). *C4.5 : Programs for Machine Learning*. Morgan Kaufmann.

RADFORD A., JÓZEFOWICZ R. & SUTSKEVER I. (2017). Learning to generate reviews and discovering sentiment. *CoRR*, **abs/1704.01444**.

SOCHER R., PERELYGIN A., WU J., CHUANG J., MANNING C. D., NG A. & POTTS C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, volume 1631, p. 1631–1642.

STUDENT (1908). The probable error of a mean. *Biometrika*, **6**(1), 1–25.

TITOV I. & MCDONALD R. T. (2008). Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th International Conference on World Wide Web, WWW 2008, Beijing, China, April 21-25, 2008*, p. 111–120.

TSYTSARAU M., PALPANAS T. & CASTELLANOS M. (2014). Dynamics of news events and social media reaction. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, p. 901–910.

TSYTSARAU M., PALPANAS T. & DENECKE K. (2010). Scalable discovery of contradictions on the web. In *Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010*, p. 1195–1196.

TSYTSARAU M., PALPANAS T. & DENECKE K. (2011). Scalable detection of sentiment-based contradictions. *DiversiWeb, WWW*, **1**, 9–16.

TURNEY P. D. (2002). Thumbs up or thumbs down ? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA.*, p. 417–424.

VOSECKY J., LEUNG K. W. & NG W. (2012). Searching for quality microblog posts : Filtering and ranking based on content analysis and implicit links. In *Database Systems for Advanced Applications - 17th International Conference, DASFAA 2012, Busan, South Korea, April 15-19, 2012, Proceedings, Part I*, p. 397–413.

WANG L. & CARDIE C. (2014). A piece of my mind : A sentiment analysis approach for online dispute detection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2*, p. 693–699.

WANG L., RAGHAVAN H., CARDIE C. & CASTELLI V. (2014). Query-focused opinion summarization for user-generated content. In *COLING 2014, 25th International Conference on Computational Linguistics, August 23-29, 2014, Dublin, Ireland*, p. 1660–1669.

YEN S.-J. & LEE Y.-S. (2006). Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset. p. 731–740.

YUAN Q., CONG G. & MAGNENAT-THALMANN N. (2012). Enhancing naive bayes with various smoothing methods for short text classification. In *Proceedings of the 21st World Wide Web Conference, WWW 2012, Lyon, France, April 16-20, 2012 (Companion Volume)*, p. 645–646.

# #AIDS Analyse Information Dangers Sexualité : caractériser les discours à propos du VIH dans les forums de santé

Yves Mercadier[1], Jérôme Azé[1], Sandra Bringay[1,2], Viviane Clavier[3], Erick Cuenca[1], Céline Paganelli[4], Pascal Poncelet[1], Arnaud Sallaberry[1,2]

[1] LIRMM, Université de Montpellier, CNRS, Montpellier, France, prenom.nom@lirmm.fr
[2] AMIS, Université Paul Valéry, Montpellier, France
[3] GRESEC EA 608 -– Université Grenoble Alpes, France
[4] LERASS EA 827, Université Paul Valéry, F34000, Montpellier, France

**Résumé** : **Contexte :** Les forums de discussion consacrés au VIH remplissent trois fonctions. Tout d'abord, ils sont utilisés par les internautes comme sources d'information pour se renseigner sur les traitements, les risques d'infection liés au sida ou le « vivre avec » la maladie ou le virus. Ensuite, ils servent de média pour les institutions de santé ou les associations afin de communiquer des informations de prévention sur le VIH. Enfin, ils apparaissent comme sources de connaissances pour les professionnels de santé (médecins, soignants) pour comprendre les comportements liés au VIH, et pour les professionnels de la prévention, pour modérer les forums et adapter les campagnes de prévention aux différents publics.
**Objectif :** Notre objectif dans cet article est de proposer un processus d'ingénierie des connaissances complet, permettant : 1) de collecter des messages dans des médias sociaux ; 2) de classifier ces messages de manière semi-automatique selon le genre, le niveau d'expertise, le niveau informationnel, le type de risque ainsi que les émotions exprimées ; 3) de visualiser ces nouvelles connaissances dans une représentation originale prenant en compte la temporalité et la hiérarchie sous-jacente à la classification. Cette visualisation pourrait permettre aux gestionnaires de sites de forums et aux professionnels de santé de naviguer dans le flot de messages pour suivre l'évolution de l'importance de ces thématiques.
**Méthodes :** Notre approche combine une démarche qualitative et quantitative. Nous avons travaillé sur plus de 226 252 messages issus du forum Sida Info Service. Deux chercheurs en sciences de l'information et de la communication ont élaboré une grille d'analyse, puis ont annoté 4 481 messages selon cette grille. Ces données ont été utilisées pour apprendre des classifieurs qui ont permis d'étiqueter l'ensemble des messages du site à notre disposition. Afin de définir les meilleurs classifieurs, nous avons comparé l'efficacité des méthodes de classification traditionnelles statistiques et plusieurs architectures d'apprentissage profond. Une fois les messages étiquetés, nous avons utilisé une visualisation de type *streamgraph*, combinée avec un outil de navigation hiérarchique, pour visualiser l'évolution de ces annotations dans le temps.
**Résultats :** Les résultats sont prometteurs et montrent l'efficacité des méthodes d'apprentissage profond pour caractériser les messages des forums de manière automatique. La méthode de visualisation mise en place permet d'explorer les résultats de ces méthodes et ainsi faciliter l'accès aux connaissances.

**Mots-clés** : Classification automatique de textes, Apprentissage profond, Visualisation, VIH

## 1 Introduction

À l'heure où 3,9 milliards d'individus dans le monde sont connectés à internet (soit 51% de la population mondiale), et où 2,91 milliards d'entre eux sont inscrits sur des réseaux sociaux (soit 29% de la population mondiale) [1], les pratiques pour s'informer et communiquer ont considérablement évolué et sont marquées par les innovations technologiques récentes et la place occupée par le numérique. En 2017, 74% des français accèdent tous les jours à Internet (95% des 18-24 ans) et passent en moyenne 1 heure 15 par jour sur les réseaux sociaux.

Les travaux présentés dans cet article se déroulent dans le cadre du projet #AIDS (Analyse Information Dangers Sexualité) qui réunit une équipe de chercheurs pluridisciplinaire en sciences de l'information et de la communication et en informatique, ainsi que des acteurs de la santé et de la prévention. Ce projet vise à analyser semi-automatiquement les contenus

1. https://www.blogdumoderateur.com/chiffres-reseaux-sociaux/

échangés sur les médias sociaux (forums de discussion, réseaux sociaux, plateformes de microblogging) consacrés au VIH pour en extraire de la connaissance. Avec plus de 35 millions de morts à ce jour, le VIH continue de représenter un problème mondial majeur de santé publique. En 2016, 1 million de personnes sont décédées d'une ou des causes liées au VIH dans le monde [2]. Toujours en 2016, en France, 6 003 personnes ont découvert leur séropositivité.

Les médias sociaux consacrés aux VIH s'adressent principalement aux personnes infectées et aux malades dans une démarche de prévention ou d'accompagnement mais également aux internautes soucieux de connaître les risques de contamination. Les publications peuvent prendre plusieurs formes comme des dispositifs de prévention, d'accompagnement, d'amélioration de la santé mentale et physique, des espaces d'échanges où sont évoqués les traitements, le "vivre avec" la maladie ou le virus, et plus largement des questions de la vie de tous les jours ou encore donner lieu à des usages sexuels d'Internet. Un grand nombre de ressources en ligne sont disponibles, à toute heure. De nombreux blogs comme *Je VIH avec* [3] alimentés par des médecins, des porteurs du virus, des associations de malades sont présents sur le web. De nombreux groupes de discussion sont consacrés à ces thématiques sur des réseaux sociaux destinés aux patients comme Carenity [4] ou encore sur le site généraliste Doctissimo [5]. On dénombre, au 1er mars 2018, une centaine de groupes facebook français en lien avec le VIH. Plus de 500 comptes twitter mentionnent les hashtags #VIH ou #SIDA dans l'intitulé de leur compte, au 1er mars 2018. Ces comptes émanent d'associations, de particuliers, d'institutions ou de médecins. Ils portent pour la plupart sur les traitements ou sur la vie quotidienne avec le VIH. Dans ce travail, nous nous focaliserons sur les messages qui nous ont été fournis par les gestionnaires du forum associé au site Sida Info Service (SIS) [6].

Si les médias sociaux ont largement été étudiés, peu de travaux ont jusqu'à présent été menés sur les médias sociaux consacrés au VIH et en particuliers les forums de santé. L'originalité de ce travail est de considérer, d'une part ces médias comme des dispositifs info-communicationnels sur lesquels se développent de nouvelles formes d'échanges et qui véhiculent des informations n'ayant pas leur place dans d'autres espaces (cabinet médical, discussions familiales, médias traditionnels...). D'autre part, les discours produits sur ces médias reflètent des pratiques ordinaires relatives aux questions de sexualité, de consommations de produits illicites et d'addictions et permettent de détecter des pratiques émergentes, de nouvelles conduites à risque. **Notre objectif dans cet article est de proposer une chaîne de traitements incluant une phase de classification et un outil de visualisation pour mieux analyser et comprendre les contenus de ces discours. Ces informations sont essentielles pour la modération du site.**

Notre approche combine une démarche qualitative et quantitative. Nous avons travaillé sur plus de $226\,252$ messages issus du forum SIS. Deux professionnels de l'information et de la communication ont élaboré une grille d'annotation, puis ont annoté $4\,481$ messages selon cette grille. Ces données ont été utilisées pour apprendre des classifieurs supervisés qui ont permis d'étiqueter dans un deuxième temps, les $226\,252$ messages de manière automatique. Une fois les messages étiquetés, nous avons utilisé une approche *MultiStream* (Cuenca *et al.*, 2018) adaptée aux séries temporelles multiples. L'originalité est ici de pouvoir visualiser simultanément la hiérarchie associée aux annotations et leur évolution dans le temps.

Concernant la phase de classification automatique de textes il s'agit d'un sujet classique de Traitement Automatique de la Langue (TAL), qui consiste à assigner des catégories prédéfinies à des documents en texte libre. Les approches de classification classiques se concentrent sur le choix du meilleur classifieur (e.g. SVM ou régression logistique) et sur la définition des meilleures caractéristiques prises en entrée de ces classifieurs. La plupart des techniques sont

2. http://www.who.int/mediacentre/factsheets/fs360/fr/
3. http://je-vih-avec.over-blog.com/
4. https://www.carenity.com/
5. http://www.doctissimo.fr/
6. https://www.sida-info-service.org/

basées sur les mots, les lexiques et sont spécifiques à une tâche particulière. Ces modèles ont été appliqués avec succès sur de très hautes caractéristiques dimensionnelles parfois éparses. Dernièrement, pour de nombreuses tâches de classification de textes, les méthodes d'apprentissage profond se sont révélées efficaces et ont permis de faire des progrès importants dans des domaines tels que la reconnaissance de formes (pattern recognition) (Baccouche *et al.*, 2011) ou la bio-informatique (Min *et al.*, 2016). Cette tendance s'est confirmée avec le succès des word embeddings (Mikolov *et al.*, 2010, 2013) et des méthodes d'apprentissage profond (Socher *et al.*, 2013). L'apprentissage profond permet l'apprentissage automatique de représentations à plusieurs niveaux. (Collobert *et al.*, 2011) ont démontré qu'une architecture d'apprentissage profond, même simple, surpasse la plupart des approches classiques pour des tâches variées de TAL telles que la reconnaissance d'entités nommées (NER) (joi, 2015), le Parsing (Vinyals *et al.*, 2015; Zhu *et al.*, 2013), l'étiquetage de rôles sémantiques (SRL) (Semantic Role Labeling) (He *et al.*, 2017; Srivastava *et al.*, 2015), le marquage POS (Andor *et al.*, 2016; Kumar *et al.*, 2016), la classification de sentiments (Rosenthal *et al.*, 2015; Nakov *et al.*, 2016; Kalchbrenner *et al.*, 2014; Kim, 2014), la traduction automatique (Sukhbaatar *et al.*, 2015). Depuis, de nombreux algorithmes complexes d'apprentissage profond ont été proposés pour résoudre ces tâches difficiles. **Dans cet article, nous allons entre autres comparer l'efficacité des méthodes de classification traditionnelles et plusieurs architectures d'apprentissage profond.**

De nombreux travaux en visualisation portent sur la représentation de données textuelles [7] (Kucher & Kerren, 2015) et/ou temporelles [8] (Aigner *et al.*, 2011). Une approche prometteuse, permettant de représenter des données temporelles issues de textes, a initialement été proposée par (Havre *et al.*, 2000) sous le nom de *ThemeRiver*. Généralisée sous le nom de *StreamGraphs*, cette approche consiste à superposer des flots représentant l'évolution d'une valeur numérique (le plus souvent issue de textes) dans le temps (Byron & Wattenberg, 2008). De nombreuses évolutions ont été proposées, e.g. (Cui *et al.*, 2011; Sun *et al.*, 2014; Wu *et al.*, 2014). La principale limite de ces approches est la difficulté de représenter un grand nombre de flots et un long intervalle de temps. La méthode utilisée dans cet article appelée *MultiStream* (Cuenca *et al.*, 2018), pallie ces problèmes en décrivant une approche de type focus+contexte combinant un ensemble de vues interactives. **Dans cet article, nous montrerons comment cette visualisation aide le chercheur en science de l'information et de la communication à analyser l'ensemble des messages du site.**

L'article est organisé en quatre sections. Dans la section 2, nous présentons notre processus. Dans la section 3, nous présentons l'évaluation de la chaîne de traitements proposée. Dans la section 4, nous discutons ces résultats et ouvrons des perspectives.

## 2 Méthodes

La figure 1 illustre la méthode proposée, structurée en 4 étapes : (1) Collecte et nettoyage de données, (2) Annotation manuelle par deux experts, (3) Prédiction automatique de catégories, (4) Visualisation temporelle et hiérarchique de ces catégories.

### 2.1 Étape 1 : Collection et nettoyage des données

Nous avons effectué une analyse de contenu de 226 252 messages en français postés sur le forum SIS qui nous ont été fournis par les gestionnaires du forum [9]. Ces messages ont été anonymisés en supprimant les pseudonymes, noms, prénoms et localités. Puis les prétraitements suivants ont été appliqués : suppression des ponctuations et des caractères spéciaux,

7. `http://textvis.lnu.se/`

8. `https://vcg.informatik.uni-rostock.de/~ct/timeviz/timeviz.html`

9. Accord CNIL, Certificat d'enregistrement 2-17031

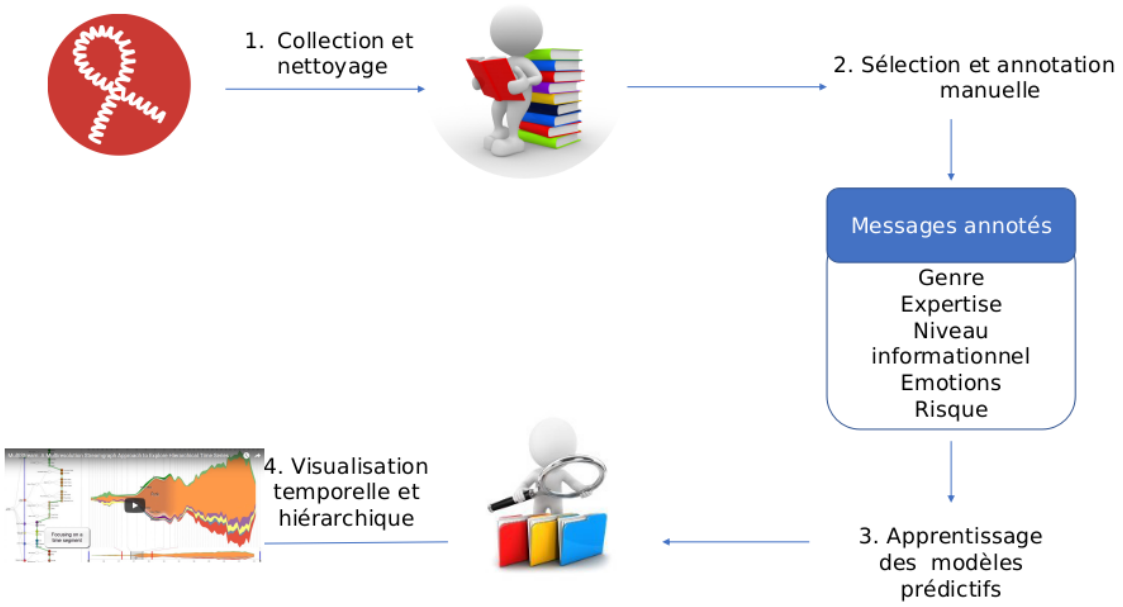



FIGURE *1 – Processus en 4 étapes*

changement des majuscules en minuscules, suppression des mots trop fréquents pour apporter de l'information à partir d'une liste de stop words et enfin stemmatisation consistant à ne conserver que le radical des mots, pour regrouper sous le même radical tous les mots d'une famille.

### 2.2 Étape 2 : Annotation manuelle

Deux chercheurs en sciences de l'information et de la communication impliqués dans le projet, ont élaboré une grille d'annotation permettant de caractériser les types d'informations véhiculées par les messages postés sur les médias sociaux à partir d'une typologie testée précédemment sur des forums de discussion de santé et en tenant compte de la temporalité de l'écriture, des thématiques abordées, des micro-échanges (Paganelli & Clavier, 2014; Clavier & Paganelli, 2017). 4 481 messages ont été annotés manuellement par les deux experts [10] (sans croisement des annotations). Les chercheurs ont systématiquement annoté l'un des forums général de SIS (le plus volumineux), qui comporte plusieurs dizaines de fils de discussion, sans faire de sélection *a priori* des thématiques. La grille s'organise selon 5 dimensions :

1. le **genre** de l'auteur décliné en femme et homme ;
2. le **niveau d'expertise** de l'auteur. On distingue les messages des modérateurs, correspondant à des savoirs, de ceux des patients, qui parfois citent leur source ou rapportent des discours ;
3. le **niveau informationnel** du message. Les messages peuvent correspondre à des partages (d'expérience ou de connaissances), à des questions formulées explicitement ou non, à des témoignages faisant référence au vécu, à l'expérience, à des informations spécialisées (qui rapportent notamment des discours professionnels) ou scientifique (contenu supposé objectif dont la validité ne s'appuie pas sur le vécu), à des conseils (informations pratiques, procédures, protocoles) ;
4. les **émotions** exprimées par les participants en considérant un spectre émotionnel : la panique, l'angoisse, la peur, l'inquiétude, la réassurance (le fait de rassurer quelqu'un) ou le fait d'être rassuré.

10. Avec l'aide de Amal Jaouzi, stagiaire, doctorante en sciences de l'information et la communication.

5. le **risque**. Un message peut faire référence à un risque encouru par les internautes lors d'un événement passé, présent ou à venir, déclencheur d'une sensation de prise de risque. Les types de situation qui déclenchent une sensation de risque n'ont pas été annotées compte tenu du fait que ces situations sont bien connues des associations et figurent dans la rubrique « Questions fréquentes ». En revanche, nous avons laissé la possibilité à l'annotateur d'introduire un commentaire libre sur l'éventuelle prise d'alcool et de drogue durant l'événement déclencheur.

Un message peut donc être décrit selon ces 5 dimensions. Dans la Table 1, nous résumons la liste des classes définies après la phase d'annotation manuelle, ainsi que le nombre de messages annotés avec ces étiquettes. Chaque message peut être associé à zéro, une ou plusieurs classes. Ces classes seront utilisées comme sorties pour la prédiction de catégories.

TABLE *1 – Sortie de la prédiction*

| **Index** | **Dimensions d'analyse** | **Classe** | **Effectif** |
|---|---|---|---|
| C1 | genre | Femme | 36 |
| C2 | genre | Homme | 67 |
| C3 | expertise | modérateur | 46 |
| C4 | expertise | savoir patient | 222 |
| C5 | expertise | sources citées | 19 |
| C6 | expertise | discours rapporté | 8 |
| C7 | information | demande | 402 |
| C8 | information | informations scientifiques ou pratiques | 661 |
| C9 | information | apport-réponse | 874 |
| C10 | information | témoignage | 284 |
| C11 | information | information vide | 1144 |
| C12 | information | conseils | 159 |
| C13 | émotion | peur | 66 |
| C14 | émotion | inquiétude | 109 |
| C15 | émotion | panique | 34 |
| C16 | émotion | général | 28 |
| C17 | émotion | angoisse | 142 |
| C18 | émotion | réassurance | 57 |
| C19 | émotion | rassuré | 22 |
| C20 | risque | anticipé | 8 |
| C21 | risque | passé | 92 |
| C22 | risque | actuel | 1 |

### 2.3 Étape 3 : Modèles pour la prédiction de catégories

Dans cette étude, nous explorons l'utilisation de méthodes d'apprentissage profond pour prédire les 22 catégories des messages collectés. Nous avons comparé les résultats des modèles d'apprentissage profond aux méthodes de classification traditionnelles. Dans cette section, nous détaillons comment chaque modèle est construit.

**Modèles.** Les modèles utilisés dans notre étude sont résumés dans le tableau 2. Les modèles d'apprentissages classiques comparés sont de type SVM (LSVC, SSVC), Naïve Bayes (MNB, BNB), Régression linéaire (SGDC), Arbre de décision (DT, AB, RF), K-plus-proches-voisins (KNC, PA). Dans ce qui suit, nous discutons plus précisément les modèles d'apprentissage profond. Tous nos modèles sont construits à partir de la même couche d'entrée et de la même couche de sortie. La couche d'entrée est une couche de type embedding permettant de caractériser chaque mot du document analysé par un vecteur de dimension 300. La couche de sortie contient elle, autant de neurones qu'il y a de classes dans notre ensemble de données. Ensuite, nous décrivons les couches intermédiaires de nos différents modèles de

réseaux neuronaux. Nous invitons le lecteur à consulter des revues plus détaillées telles que celle de (Schmidhuber, 2015).

TABLE *2 – Modèles de Classification : classiques vs apprentissage profond*

| Classiques | Apprentissage profond |
|---|---|
| LinearSVC (LSVC)<br>SigmoïdSVC (SSVC)<br>MultinomialNB (MNB)<br>BernoulliNB (BNB)<br>SGDClassifier (SGDC)<br>Decision Tree (DT)<br>AdaBoost (AB)<br>Random Forest (RF)<br>KNeighborsClassifier (KNC) | Basic neural network (BNN)<br>Multi-layer perceptrons (MLP)<br>Deep Multi-layer perceptrons (DMLP)<br>Long short-term memory (LSTM) RNN<br>Bidirectional Long short-term memory (BLSTM) RNN<br>Gated Recurrent Unit (GRU) RNN<br>Long-term Recurrent Convolutional Networks (LRCN)<br>Convolutional neural network (CNN)<br>Serial Convolutional neural network (SCNN)<br>Multi Channel Convolutional neural network (MCNN) |

— *Basic neural network (BNN)* connecte directement la couche d'entrée avec la couche de sortie.
— *Multilayer Perceptrons (MLP)* est le modèle le plus simple d'apprentissage profond. Il est réalisé par une couche cachée de neurones entre la couche d'entrée et de sortie.
— *Deep Multilayer Perceptrons (DMLP)* est basé sur 20 couches cachées.
— *Long Short-Term Memory (LSTM)* est un sous-type de RNN. Les réseaux neuronaux récurrents permettent l'étude de séquences de données. Ces réseaux, composés de plusieurs couches, estiment leurs sorties en fonction des états de leurs couches précédentes en utilisant une mémoire intermédiaire. LSTM est basé sur des blocs de mémoire qui sont utilisés comme unités dans la couche récurrente pour capturer des dépendances à plus longue portée.
— *Bidirectional Long Short-Term Memory (BLSTM)* permet d'utiliser l'information après et avant les données étudiées par le réseau au temps $t$.
— *Gated Recurrent Unit (GRU)* est un réseau de neurones permettant une étude plus rapide que le LSTM tout en conservant les qualités de celui-ci.
— *Long-term Recurrent Convolutional Networks (LRCN)* associe en série un réseau convolutif et un réseau LSTM.
— *Convolutional Neural Network (CNN)* sont structurés par deux opérations : convolution puis max-pooling. La convolution est basée sur plusieurs filtres combinés pour extraire les nombreuses propriétés associées aux données. La seconde opération compresse les résultats de l'opération précédente pour extraire des informations denses.
— *Serial Convolutional neural network (SCNN)* place séquentiellement quatre réseaux convolutionnels les uns après les autres.
— *Multi Channel Convolutional neural network (MCNN)* est basé sur six réseaux de convolution branchés sur la couche d'entrée. Chaque réseau de convolution possède un noyau différent des autres. Ensuite, nous associons la sortie des six réseaux de convolution à une seule couche pour construire une sortie mono-couche entièrement connectée.

**Entrées des modèles :** Les entrées des modèles sont les messages. Nous utilisons deux pré-traitements distincts pour les deux types de méthodes de classification étudiées. Pour les méthodes de classification traditionnelles, nous appliquons une vectorisation basée sur la mesure Tf-Idf (Ramos, 1999). Pour les méthodes d'apprentissage profond, les données sont préparées différemment. Nous conservons les $5000$ mots les plus fréquents (hors stop words) et nous représentons chaque document par une séquence de mots ce qui permet de conserver l'ordre des mots.

**Sorties des Modèles** : Les effectifs dans les classes sont très déséquilibrés. Nous avons donc mis en place deux stratégies que nous appelons *précise* et *imprécise*. Dans la straté-

gie *précise*, l'objectif est de décrire finement un message, par exemple selon la dimension émotion, en l'associant à l'étiquette *peur* ou *angoisse*. Dans la stratégie *imprécise*, le niveau de description attendu est plus général et le message pourrait être associé à une *émotion négative*. Nous avons donc regroupé manuellement, après discussion avec les experts, certaines classes. Par ailleurs, un grand nombre de messages ne sont pas étiquetés comme le montre le tableau 1. Pour le niveau d'expertise, les émotions et le risque, nous avons considéré une classe *Autres* correspondant à des messages non étiquetés. Nous effectuons sur ces échantillons un ré-échantillonnage de la taille de la classe majoritaire (non étiquetée). Une description des classes finales est donnée dans la table 3.

TABLE *3 – Classes utilisées dans les modèles selon les deux stratégies précise et imprécise*

| **Stratégie** | **Précise (Nombre de posts)** | **Imprécise (Nombre de posts)** |
|---|---|---|
| **Genre** | *genre_1*<br>C1 (57)<br>C2 (28) | |
| **Niveau d'expertise** | *expe_1*<br>C3 (222)<br>C4 (46)<br>C5/C6 (74) | *expe_2*<br>C3/C4/C5/C6 (342)<br>Autres (342) |
| **Niveau informationnel** | *info_1*<br>C7 (402)<br>C8 (661)<br>C9 (874)<br>C10 (284)<br>C11 (1144)<br>C12 (159) | *info_2*<br>C7 (402)<br>C8/C9/C10/C12 (1978)<br>C11 (1144) |
| **Émotion** | *emot_1*<br>C13/C14 (65)<br>C15/C17 (544)<br>C18/C19 (30) | *emot_2*<br>C13/C14/C15/C16 (609)<br>C18/C19 (30)<br>Autres (609) |
| **Risque** | | *risq*<br>C20/C21/C22 (101)<br>Autres (101) |

**Partitionnement des données et entraînement** Afin d'évaluer les différents modèles, nous appliquons une validation croisée en dix plis. L'ensemble de données est divisé en dix sous-ensembles. Nous utilisons neuf sous-ensembles pour la phase d'apprentissage et un sous-ensemble pour la phase de validation. Nous avons répété ce processus dix fois. Pour chaque pli, nous utilisons un sous-ensemble de différentes phases de validation et nous avons calculé une métrique pour évaluer la performance.

Pour les algorithmes de classification classiques, nous utilisons l'outil sklearn [1] avec les paramètres par défaut. Pour toutes les architectures d'apprentissage profond, nous avons utilisé les paramètres par défaut pour la taille du Mini-batch (c.-à-d. Le nombre d'instances d'entraînement à considérer en même temps), la dimension d'intégration (embedding dimension, chaque mot est décrit par un vecteur de dimension $n$), le nombre d'epoch (le nombre d'itérations sur l'ensemble d'entraînement), la fonction d'activation et le taux d'abandon (Dropout ratio, ratio d'unités cachées à désactiver dans chaque formation de Mini-batch). Ensuite, nous utilisons une configuration avec un Mini-batch de taille 32, une dimension d'intégration de 300, une couche cachée de taille 256, un nombre d'epoch de 25, la fonction d'activation selu et un Dropout de 0,3.

1. `http://scikit-learn.org/stable/`

**Mesure de la performance** Pour tous les modèles, nous nous sommes basés sur une métrique couramment utilisée dans le domaine de la classification. Chaque sortie est un vecteur de dimension $n$ correspondant aux $n$ classes à prédire. Nous mesurons la qualité des classifieurs avec la métrique *exactitude* (accuracy). Celle-ci exprime le nombre de classes prédites justes en regard du nombre total de prédictions de classes réalisées. Nous avons choisi d'utiliser une métrique stricte en considérant une sortie comme juste uniquement si l'intégralité des classes prédites sont exactes.

**Exemple 1**
*Considérons 3 documents décrits par un vecteur d'annotations pour 4 classes. Chaque élément $i$ de ce vecteur correspond à la présence (1) ou à l'absence (0) de la classe $i$ pour le document.*
$Document1 \longrightarrow 0010$
$Document2 \longrightarrow 1000$
$Document3 \longrightarrow 0100$
*Considérons maintenant un modèle retournant 3 vecteurs de prédictions de ces 4 classes. Chaque élément $i$ de ce vecteur correspond à la prédiction (1) ou à l'absence de prédiction (0) de la classe $i$ pour le document.*
$prédiction1 \longrightarrow 0010$
$prédiction2 \longrightarrow 0010$
$prédiction3 \longrightarrow 1100$
*Seule la première prédiction sera considérée comme juste pour les trois prédictions émises. L'exactitude du modèle sera égale à : $acc = \frac{1}{3} = 0.333$*

### 2.4 Étape 4 : Visualisation temporelle et hiérarchique

Le flux de messages peut être représenté sous la forme d'une série temporelle multiple, c'est-à-dire un ensemble de variables quantitatives associées à un même intervalle temporel. Au cours des dernières années, les visualisations de type *Streamgraphs* (Byron & Wattenberg, 2008) ont été largement utilisées pour représenter l'évolution de plusieurs séries temporelles. Cependant, les *Streamgraphs* posent deux problèmes majeurs : (1) Comment gérer un long intervalle de temps ? (2) Comment gérer un grand nombre de séries temporelles ?

Pour résoudre le problème (1), des méthodes de type vue d'ensemble+détail ou encore focus+contexte peuvent être mises en place (voir (Munzner, 2014) pour une introduction). Les méthodes vue d'ensemble+détail consistent à combiner deux vues, une montrant l'ensemble des données de façon agrégée, l'autre montrant en détail un sous ensemble sélectionné. Les méthodes de type focus+contexte consistent à intégrer les parties agrégées (contexte) et les parties détaillées (focus) dans une seule et même vue. La plus connue de ces méthodes est sans doute le *fisheye* qui consiste à appliquer une distortion sur la visualisation de façon à agrandir la zone du focus en réduisant la zone du contexte.

Pour résoudre le problème (2), les séries temporelles peuvent être aggrégées selon leur proximité pour former une structure hiérarchique visualisable sous forme d'arbre. Cet arbre peut aider à sélectionner les niveaux de détail affichés.

La méthode que nous utilisons ici (Cuenca *et al.*, 2018) repose sur 3 vues permettant de traiter les problèmes susmentionés. Une première vue (voir Fig. 2 cadre 1) permet de voir les séries agrégées sur l'ensemble de l'intervalle de temps, et de sélectionner un sous-intervalle (vue d'ensemble). Une seconde vue (voir Fig. 2 cadre 2) affiche les séries sur l'intervalle sélectionné (détail). Cette vue inclut un *fisheye*, i.e. une zone dans laquelle les séries sont détaillées (focus), entourée par des zones dans lesquelles les séries sont agrégées (contexte). La troisième vue (voir Fig. 2 cadre 3) permet de visualiser l'arbre d'agrégation, dans notre cas la hiérarchie des classes présentées dans la table 1 et de sélectionner les niveaux devant être utilisés pour les zones agrégées et pour les zones détaillées.


FIGURE 2 – *Visualisation des séries temporelles avec un focus sur la période 2008-2011.*


FIGURE 3 – *Visualisation de la dimension "Niveau informationnel" dans laquelle les messages relatifs à la classe "Information vide" ont été filtrés grâce à l'arbre d'agrégation.*

## 3 Résultats

### 3.1 Comparaison des modèles de classification

TABLE *4 – Comparaison des performances des classifieurs.*

| Stratégie | Précise | | | | Imprécise | | | |
|---|---|---|---|---|---|---|---|---|
| jeux | emot_1 | genre_1 | expe_1 | info_1 | emot_2 | expe_2 | info_2 | risq |
| LSVC | 0.318 | 0.621 | 0.735 | 0.255 | 0.533 | 0.659 | 0.488 | **0.638** |
| SSVC | **0.325** | 0.624 | 0.735 | 0.242 | 0.532 | 0,422 | 0,000 | 0,440 |
| MNB | 0,020 | 0.631 | 0.701 | 0.175 | 0.472 | 0.662 | 0.406 | 0.603 |
| BNB | 0.058 | **0.660** | 0.725 | 0.280 | **0.563** | **0.672** | 0.371 | 0.603 |
| SGDC | 0.309 | 0.640 | 0.711 | 0.238 | 0,518 | 0.630 | **0.490** | 0.608 |
| DT | 0.244 | 0.533 | 0.664 | 0.206 | 0.468 | 0.625 | 0.362 | 0.564 |
| AB | 0.300 | 0.563 | 0.718 | 0.248 | 0.493 | 0.600 | 0.450 | 0.633 |
| RF | 0.0 | 0.650 | **0.752** | 0.0 | 0.116 | 0.379 | 0.0 | 0.440 |
| KNC | 0.227 | 0.533 | 0.677 | **0.323** | 0.439 | 0.589 | 0.281 | 0.461 |
| BNN | 0.718 | 0.945 | 0.966 | 0.372 | 0.841 | 0.894 | 0.883 | 0.883 |
| MLP | 0.662 | 0.945 | 0.966 | 0.356 | 0.745 | 0.901 | 0.909 | **0.921** |
| DMLP | 0.627 | 0.658 | 0.788 | 0.308 | 0.801 | 0.342 | 0.293 | 0.376 |
| LSTM | **0.827** | **0.990** | **0.969** | **0.564** | **0.882** | **0.922** | **0.940** | 0.906 |
| BLSTM | 0.776 | 0.927 | 0.953 | 0.521 | 0.855 | 0.906 | 0.898 | 0.902 |
| GRU | 0.502 | 0.684 | 0.884 | 0.355 | 0.441 | 0.805 | 0.745 | 0.805 |
| LRCN | 0.795 | 0.954 | 0.926 | 0.516 | 0.832 | 0.881 | 0.893 | 0.907 |
| CNN | 0.713 | 0.954 | 0.963 | 0.408 | 0.841 | 0.900 | 0.883 | 0.897 |
| SCNN | 0.741 | 0.954 | 0.959 | 0.436 | 0.829 | 0.905 | 0.893 | 0.892 |
| MCNN | 0.734 | 0.963 | 0.956 | 0.382 | 0.847 | 0.905 | 0.881 | 0.892 |

Le tableau 4 synthétise les résultats des différents classifieurs sur les différentes dimensions et selon les 2 stratégies. Comme l'on pouvait s'y attendre, les résultats pour les stratégies de classification précises sont inférieurs aux résultats pour les stratégies de classification imprécises. Par exemple, pour la dimension relative au niveau informationnel, l'exactitude pour la classification info_1 (stratégie précise) est au mieux de 0,564 alors que pour info_2 (stratégie imprécise) est de 0,906. Pour chaque type de classification, les meilleurs classifieurs traditionnels ont obtenu des résultats inférieurs aux classifieurs en apprentissage profond. Par exemple, pour la classification du genre, le meilleur classificateur traditionnel obtient une exactitude de 0,660 alors que le meilleur classifieur en apprentissage profond obtient une exactitude de 0,990. Pour 7 types de classification sur 8, la meilleure architecture est LSTM. Ce résultat est cohérant avec la littérature. Par exemple, Johnson & Zhang (2016) obtiennent également des résultats performants sur des textes avec LSTM. La classification du genre, de l'expertise, du risque et du niveau d'expérience obtiennent en apprentissage profond des exactitudes supérieures à 0,9 dans le cas imprécis et même parfois précis. Les résultats sont plus mitigés pour les émotions et le niveau informationnel qui correspondent aux classes très déséquilibrées, en particulier avec les classifieurs classiques. Par exemple, le classifieur RF ne fonctionne pas sur les classes déséquilibrées telles que emot_1 ou emot_2. Il serait intéressant de réannoter un nombre suffisant d'exemples pour rééquilibrer ces classes.

### 3.2 Visualisation

La visualisation obtenue à partir des $226\,252$ messages annotés automatiquement à partir du meilleur modèle (LSTM) est disponible sur le site du projet Multistream [11]. La figure 2 permet de repérer un pic d'activité du forum en 2009 et une majorité d'information de type

11. `http://advanse.lirmm.fr/multistream/` - jeu de données AIDS

"Information vide" révélant la part importante de messages de convivialité ou de soutien, les conseils étant également très recherchés dans des forums modérés par des professionnels de la prévention. La figure 3 est obtenue après sélection des classes à afficher dans l'arbre d'agrégation, en filtrant les messages de type "Information vide". On remarque alors que l'information de type "Demande" domine le pic de messages de 2009, ce qu'il n'était pas possible de visualiser avant le filtrage. Ces deux figures illustrent le processus de navigation dynamique dans le flux de messages, rendu possible par l'outil *MultiStream*.

Nous allons détailler dans la suite les 5 dimensions identifiées lors de la phase d'annotation manuelle (le genre, le niveau d'expertise, le niveau informationnel, le risque et les émotions) et illustrer avec des exemples ce qu'il est possible de repérer dans ces flots d'informations, et l'exploitation éventuelle qui peut en être faite.

Concernant la dimension *Genre*, on remarque qu'il y a plus de messages d'hommes que de femmes. Cette situation est spécifique au VIH, puisque les femmes sont souvent plus présentes sur les forums de santé (Paganelli & Clavier, 2011). Dans le cas du sida, les hommes ayant des relations sexuelles avec d'autres hommes (HARSAH), en termes de prévalence, sont les plus touchés par le VIH, ils sont aussi la cible des campagnes de prévention et sont bien informés des risques d'exposition au sida (De Oliveira, 2012). Ces résultats nous laissent penser que les forums constituent des médias alternatifs et complémentaires (Renahy & Chauvin, 2006) indispensables pour des hommes plutôt moins acculturés aux messages de prévention, qui se renseignent pour des prises de risques plus exceptionnelles (par exemple dans le cadre de relations extra-conjugales hétéro ou homosexuelles).

Concernant la dimension *Niveau d'expertise*, les messages annotés montrent que se développent sur les forums de nouvelles formes d'expertise, hors de la sphère médicale et relayées par le biais des participants ou des modérateurs. Ces informations sont de première importance pour les professionnels (de santé et de la prévention) qui peuvent ainsi appréhender les connaissances qui circulent sur le VIH. Ainsi, la majorité des messages correspondent à des "savoirs patients" et dans une moindre proportion à des messages de modérateurs ou à des messages contenant des sources citées et des discours rapportés. Dans le cas du VIH, les patients apparaissent comme experts de leur propre maladie, de leur cas singulier bien qu'ils ne soient pas spécialistes de la maladie en général ni de la médecine. Ces derniers ont été conduits à côtoyer des professionnels de santé dont ils rapportent les propos, ce qui légitime les conseils qu'ils donnent aux autres participants : *"Merci de m'avoir répondu, c'est vrai vous m'avez bien rassurée ms je ne peux m'empecher de penser qu'il y a un risque , j'en ai parlé à 2 médecins qui travaillent avec moi , ils me disent que le risque est minime ms qu'il y en a un , ils ont l'habitude de soigner des séropositifs puisqu'ils accueillent des patients ayant eu un accident exposant au sang comme on dit chez nous"*. Les forums sont aussi le moyen de transmettre des sources spécialisées qui participent à la circulation des savoirs sur le VIH : *"En tout cas pour tout ce qui est des modes de transmission du vih tu trouveras tout ce qu'il te faut savoir ici : http ://www.sida-info service.org/forums ... hp4 ?t=3257[ . . . ]."*.

Concernant le *Niveau informationnel*, la majorité des messages sont vides, au sens où ils renvoient à des messages de convivialité (ex. *"Bonjour tous", "@ bientôt", "Merci++"*). Une fois ces messages filtrés, deux types d'informations apparaissent : des demandes ou des apports d'information associés à ces demandes. Ces messages concernent essentiellement des informations pratiques ou scientifiques : *"Bonjour, J'ai lu que les anticorps apparaissent generalement entre 20 et 45 jours apres un risque. Ca veut donc dire que ||nom|| n'est tres fiable qu'apres 6 semaines ? Je sens que je vais avoir la reponse sybilline "1 mois=bonne indication, 3 mois= sur", mais j'aimerais savoir ce que ca veut dire bonne indication (je suis negatif a 4 semaines avec des ganglions et des courbatures partout... Et pouvezvous aussi me dire si le test ||nom|| detecte le VIH-1 et le VIH-2 ? Merci pour vos reponses ! !"*, et dans une moindre mesure des témoignages relatant le vécu des porteurs du virus. Les participants au forum directement concernés par le VIH cherchent en premier lieu des informations relatives à cette maladie, aux symptômes, aux traitements. Ils cherchent également à mieux comprendre le discours des experts en amont d'un rendez-vous médical ou en aval pour trouver la définition

de termes spécialisés ou confirmer le discours des professionnels de santé. Pour les proches non directement concernés par le virus, la recherche d'information permet de découvrir des conseils pratiques pour accompagner une personne malade.

Concernant le *risque*, il faut distinguer une dimension objective du risque, incarnée dans une connaissance plus ou moins partagée des facteurs de prise de risque, qui se confronte à des perceptions variables du risque en lien notamment avec l'expérience personnelle de l'individu, son environnement social, affectif et culturel. La dimension subjective du risque motive l'essentiel des demandes : *"Bonjour, ma copine est séropositive depuis peu, mon 1er test est négatif. J'attends 3 mois pour confirmer le résultat. Comme j'entends tout et n'importe quoi a propos des fellations, je voulais savoir si je risquais un risque ou pas lors d'une fellation dans mon cas. Merci pour votre réponse. C'est nouveau pour moi et je gère assez mal la situation."* Ainsi, les trois quarts des messages portent sur un risque passé, peu de personnes posant des questions avant une situation présumée risquée. Dans les forums où priment l'émotion, c'est la panique, le stress qui dominent. Cette situation engendre de la part des répondants des messages de réassurance, la catégorie d'émotion qui est la plus représentée en termes de pourcentage d'apports d'information. En effet, de nombreux messages d'encouragements, par exemple lors de l'attente d'un dépistage viennent soutenir les membres ayant pris un risque. Peu de messages ont cependant été étiquetés selon le risque lié à l'alcool et à la prise de substance psychotrope. Ce constat est certainement lié au fait qu'un forum institutionnel de type SIS n'est pas le lieu pour évoquer les pratiques à risques autres que celles liées à la sexualité.

### 3.3 Limites de l'étude

Notre approche préliminaire présente de nombreuses limites que nous détaillons dans cette section.

Concernant la phase d'annotation, la grille utilisée pourrait être retravaillée. La dimension relative au *Niveau Informationnel* pourrait être scindée en deux pour distinguer les messages relevant de l'interaction (demande ou apport de réponse à une question) des messages uniquement destinés à l'apport d'information (Information Scientifique et pratique, témoignage et conseil) ou encore les messages vides au niveau informationnel. Cela est tout à fait possible car la visualisation utilisée permet d'afficher plusieurs niveaux de hiérarchie.

Concernant la phase de classification, la principale limite est que nous avons utilisé uniquement l'exactitude comme métrique pour évaluer la performance des classifieurs. Le rappel et la précision sont deux mesures importantes pour évaluer la qualité d'un classifieur et devrait être prises en considération. Une autre limitation porte sur l'interprétation de l'apprentissage profond. Ces modèles sont des "boîtes noires". Ils ne fournissent pas d'explication même si la prédiction est efficace (Shwartz-Ziv & Tishby, 2017). Cependant, malgré un nombre de messages manuellement annotés limité, cette étude a montré que la performance des modèles traditionnels était beaucoup plus faible que la performance de l'apprentissage profond. De nouvelles techniques d'apprentissage profond sont étudiées pour faciliter l'interprétation de tels modèles (Liu *et al.*, 2017; Lipton, 2016) et pourraient s'avérer judicieuses dans ce contexte.

Nous avons également identifié des limites liées à la temporalité utilisée pour la visualisation de l'ensemble des données du site SIS. Actuellement, la temporalité est le mois. Or, la période temporelle d'interaction est généralement le jour ou la semaine dans le contexte des échanges sur les forums. Une possibilité de focus sur une période plus courte permettrait de repérer des épisodes liés à des questions sur une prise de risque et le délai pour se faire dépister. En effet, 48h après une prise de risque, il existe un traitement préventif qui peut être suggéré par les internautes. Les résultats du premier test permettant de savoir s'il y a eu contamination sont obtenus entre 6 semaines et 90 jours. Cette période s'accompagne de messages contenant des émotions négatives, de la réassurance et dans la majorité des cas se finit par des messages de la personne rassurée comme nous avons pu l'observer dans de nombreux fils de discussion. De même, la visualisation pourrait être améliorée par l'affichage d'informations supplémentaires comme l'heure des messages. Cette information est essentielle notamment

lorsque l'on s'intéresse aux émotions exprimées dans les messages. Par exemple, les angoisses s'expriment généralement la nuit. La quantification de ces phénomènes est important pour le site SIS en terme de planification de leurs équipes de modérateurs.

Pour finir, une dernière limite de notre étude est liée à la généralisation de nos résultats. En effet, la tâche est très spécifique au sujet d'étude, le VIH, au type de textes que sont les messages des forums, ce qui rend difficile la généralisation de notre approche. Toutefois, notre méthodologie et les résultats peuvent être utilisés comme référence pour d'autres études sur l'identification automatique de catégories à partir de données sociales.

## 4 Conclusion et perspectives

Notre étude a souligné l'efficacité des architectures d'apprentissage profond pour prédire des thèmes sur des données de forum. La visualisation originale sous forme de StreamGraph permet d'explorer de manière efficace tout en quantifiant les différentes catégories de messages. La faisabilité de notre approche peut conduire à de nouvelles applications en santé basées sur les médias sociaux destinés aux patients et aux professionnels de santé.

Nos résultats montrent clairement le besoin d'études capables d'analyser automatiquement des forums et d'en extraire des informations utiles. Nous proposons l'utilisation de l'apprentissage automatique et de la visualisation interactive pour relever ces défis. Cette étude reste préliminaire. Une étude plus approfondie sur les différents types de préparation des données, paramètres des algorithmes, modèles d'apprentissage permettrait d'affiner l'interprétation des résultats de la phase 3. En particulier, différents regroupement de classes et l'interprétation des liens entre classes serait pertinent. Quels sont les sentiments associés à une prise de risque ? Est-ce que les réponses apportées dans les fils sont rassurantes ou satisfaisantes ? Comment s'informent les participants dans les forums ? Pour finir, une étude poussée sur l'utilité et l'utilisabilité de la visualisation présentée en phase 4 est également nécessaire. Nous suggérons 4 perspectives.

Premièrement, nous prévoyons d'entreprendre une analyse à grande échelle en utilisant une collection de médias sociaux plus large (Autres forums, Facebook, Twitter, ...). Cette analyse inclura l'application de méthode d'apprentissage non supervisé de type Latent Dirichlet Allocation (LDA) (Pennacchiotti & Gurumurthy, 2011; Wang *et al.*, 2012) pour extraire les thèmes émergeant des discussions et l'exploration du style linguistique des différents utilisateurs (Zhan *et al.*, 2017; Zeng & Tse, 2006; Wang *et al.*, 2014). Une attention particulière sera portée sur l'identification d'une typologie des risques encourus en lien avec le VIH (e.g. consommation de drogue, comportements sexuels atypiques, etc).

Deuxièmement, nous pensons que lorsque les ensembles de données sont petits, l'apprentissage est difficile. Une amélioration significative serait la mise en œuvre de techniques d'apprentissage actif (Olsson, 2009). En effet, dans ce type de tâche, il est important d'optimiser les informations disponibles afin que les systèmes de classification puissent les utiliser le plus efficacement possible pendant la phase d'apprentissage tout en préservant l'acquisition de nouveaux échantillons étiquetés (Ducoffe & Precioso, 2015). L'utilisation de la visualisation pour guider les annotateurs vers des messages à annoter pourrait s'avérer également intéressante.

Troisièmement, au sein d'un ensemble de données suffisamment important, nous pouvons tirer parti des modèles d'apprentissage automatique pour utiliser des fonctionnalités plus complexes pour caractériser les utilisateurs qui postent ces messages. Nous suggérons de mettre l'accent sur les groupes d'utilisateurs, y compris les professionnels de la santé, les célébrités, le grand public et les associations. Cela nous amènera à comprendre quel groupe d'utilisateurs est important, peut jouer le rôle d'influenceur, les incitant à partager leurs messages, à les aimer et à leur répondre.

Pour finir, nous prévoyons d'étudier la distribution temporelle des messages pour nous concentrer sur la dynamique des thématiques au fils du temps. Nous pouvons étudier les corrélations temporelles entre les réactions des internautes et les événements du monde réel comme les soirées de type Sidaction. Cette analyse exploratoire pourrait aider à identifier les

facteurs contribuant à la sensibilisation. Au-delà, nous pouvons également analyser la répartition géographique des messages.

Ce type d'étude est importante pour convaincre les parties prenantes, les professionnels de la santé et le grand public de s'impliquer et d'utiliser le Web 3.0 comme intelligence collective pour repousser les maladies telles que le VIH.

## 5 Remerciements

Ce travail s'intègre dans le projet #AIDS et a été soutenu par une subvention ANRS [12] en 2016. Les auteurs souhaitent remercier les gestionnaires du site Sida-Info-Service pour le partage des données et ses participants pour leur engagement à combattre le VIH.

## Références

(2015). *Joint Named Entity Recognition and Disambiguation*.

AIGNER W., MIKSCH S., SCHUMANN H. & TOMINSKI C. (2011). *Visualization of Time-Oriented Data*. Springer.

ANDOR D., ALBERTI C., WEISS D., SEVERYN A., PRESTA A., GANCHEV K., PETROV S. & COLLINS M. (2016). Globally normalized transition-based neural networks. cite arxiv :1603.06042.

BACCOUCHE M., MAMALET F., WOLF C., GARCIA C. & BASKURT A. (2011). Sequential deep learning for human action recognition. In *Proceedings of the Second International Conference on Human Behavior Unterstanding*, HBU'11, p. 29–39, Berlin, Heidelberg : Springer-Verlag.

BYRON L. & WATTENBERG M. (2008). Stacked Graphs - Geometry & Aesthetics. *IEEE Transactions on Visualization and Computer Graphics*, **14**(6), 1245–1252.

CLAVIER V. & PAGANELLI C. (2017). Une approche méthodologique croisée du traitement des données de la recherche : le cas d'un corpus d'échanges issus de médias sociaux dans le domaine de la santé. In *Colloque COSSI 2017, Méthodes et stratégies de gestion de l'information par les organisations : des "Big Data"aux "Thick Data", 85ème congrès de l'ACFAS, Université McGill*.

COLLOBERT R., WESTON J., BOTTOU L., KARLEN M., KAVUKCUOGLU K. & KUKSA P. (2011). Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, **12**, 2493–2537.

CUENCA E., SALLABERRY A., WANG F. Y. & PONCELET P. (2018). MultiStream : A Multiresolution Streamgraph Approach to Explore Hierarchical Time Series. *IEEE Transactions on Visualization and Computer Graphics*, (to appear).

CUI W., LIU S., TAN L., SHI C., SONG Y., GAO Z., QU H. & TONG X. (2011). TextFlow : Towards Better Understanding of Evolving Topics in Text. *IEEE Transactions on Visualization and Computer Graphics*, **17**(12), 2412–2421.

DE OLIVEIRA J.-P. (2012). *Communication publique et formes de gouvernabilité contemporaines de l'Etat : le cas de l'homosexualité dans les campagnes de prévention du sida en France (1987-2007)*. PhD thesis, Université Stendhal, Grenoble.

DUCOFFE M. & PRECIOSO F. (2015). QBDC : query by dropout committee for training deep supervised architecture. *CoRR*, **abs/1511.06412**.

HAVRE S., HETZLER E. & NOWELL L. (2000). ThemeRiver : Visualizing Theme Changes over Time. In *Proceedings of the IEEE Symposium on Information Visualization*, p. 115–123 : IEEE.

HE L., LEE K., LEWIS M. & ZETTLEMOYER L. (2017). Deep semantic role labeling : What works and what's next. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

JOHNSON R. & ZHANG T. (2016). Supervised and semi-supervised text categorization using lstm for region embeddings. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, p. 526–534 : JMLR.org.

KALCHBRENNER N., GREFENSTETTE E. & BLUNSOM P. (2014). A convolutional neural network for modelling sentences. In *ACL (1)*, p. 655–665 : The Association for Computer Linguistics.

KIM Y. (2014). Convolutional neural networks for sentence classification. In A. MOSCHITTI, B. PANG & W. DAELEMANS, Eds., *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, p. 1746–1751 : ACL.

12. http://www.anrs.fr/fr

KUCHER K. & KERREN A. (2015). Text visualization techniques : Taxonomy, visual survey, and community insights. In *Proceedings of the IEEE Pacific Visualization Symposium (PacificVis)*, p. 117–121.

KUMAR A., IRSOY O., ONDRUSKA P., IYYER M., BRADBURY J., GULRAJANI I., ZHONG V., PAULUS R. & SOCHER R. (2016). Ask me anything : Dynamic memory networks for natural language processing. In M. F. BALCAN & K. Q. WEINBERGER, Eds., *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, p. 1378–1387, New York, New York, USA : PMLR.

LIPTON Z. C. (2016). The mythos of model interpretability. *CoRR*, **abs/1606.03490**.

LIU W., WANG Z., LIU X., ZENG N., LIU Y. & ALSAADI F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing*, **234**, 11 – 26.

MIKOLOV T., KARAFIÁT M., BURGET L., CERNOCKÝ J. & KHUDANPUR S. (2010). Recurrent neural network based language model. In T. KOBAYASHI, K. HIROSE & S. NAKAMURA, Eds., *INTERSPEECH*, p. 1045–1048 : ISCA.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S. & DEAN J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. BURGES, L. BOTTOU, M. WELLING, Z. GHAHRAMANI & K. Q. WEINBERGER, Eds., *Advances in Neural Information Processing Systems 26*, p. 3111–3119. Curran Associates, Inc.

MIN S., LEE B. & YOON S. (2016). Deep learning in bioinformatics. *CoRR*, **abs/1603.06430**.

MUNZNER T. (2014). *Visualization Analysis and Design*. A.K. Peters visualization series. A K Peters.

NAKOV P., RITTER A., ROSENTHAL S., SEBASTIANI F. & STOYANOV V. (2016). Semeval-2016 task 4 : Sentiment analysis in twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, p. 1–18.

OLSSON F. (2009). *A literature survey of active machine learning in the context of natural language processing*. Rapport interne T2009 :06.

PAGANELLI C. & CLAVIER V. (2011). Le forum de discussion : une ressource informationnelle hybride entre information grand public et information spécialisée. *Yasri-Labrique Eleonore. Les forums de discussion : agoras du XXIe siècle ? Théories, enjeux et pratiques discursives, L'harmattan (collection Langue et Parole)*, p. 39–55.

PAGANELLI C. & CLAVIER V. (2014). S'informer via des médias sociaux de santé : quelle place pour les experts ? **23**, 141–143.

PENNACCHIOTTI M. & GURUMURTHY S. (2011). Investigating topic models for social media user recommendation. In *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, p. 101–102, New York, NY, USA : ACM.

RAMOS J. (1999). Using tf-idf to determine word relevance in document queries.

RENAHY E. & CHAUVIN P. (2006). Internet uses for health information seeking : A literature review. *Revue Epidémiologique de Santé Publiqu*, **54**(3), 263–275.

ROSENTHAL S., NAKOV P., KIRITCHENKO S., MOHAMMAD S., RITTER A. & STOYANOV V. (2015). Semeval-2015 task 10 : Sentiment analysis in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, p. 451–463.

SCHMIDHUBER J. (2015). Deep learning in neural networks : An overview. *Neural Networks*, **61**, 85–117. Published online 2014 ; based on TR arXiv :1404.7828 [cs.NE].

SHWARTZ-ZIV R. & TISHBY N. (2017). Opening the black box of deep neural networks via information. *CoRR*, **abs/1703.00810**.

SOCHER R., PERELYGIN A., WU J. Y., CHUANG J., MANNING C. D., NG A. Y. & POTTS C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, volume 1631, p. 1642.

SRIVASTAVA R. K., GREFF K. & SCHMIDHUBER J. (2015). Training very deep networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, p. 2377–2385, Cambridge, MA, USA : MIT Press.

SUKHBAATAR S., SZLAM A., WESTON J. & FERGUS R. (2015). End-to-end memory networks. In C. CORTES, N. D. LAWRENCE, D. D. LEE, M. SUGIYAMA & R. GARNETT, Eds., *Advances in Neural Information Processing Systems 28*, p. 2440–2448. Curran Associates, Inc.

SUN G., WU Y., LIU S., PENG T.-Q., ZHU J. J. H. & LIANG R. (2014). EvoRiver : Visual Analysis of Topic Coopetition on Social Media. *IEEE Transactions on Visualization and Computer Graphics*, **20**(12), 1753–1762.

VINYALS O., KAISER L. U., KOO T., PETROV S., SUTSKEVER I. & HINTON G. (2015). Grammar as a foreign language. In C. CORTES, N. D. LAWRENCE, D. D. LEE, M. SUGIYAMA & R. GARNETT, Eds., *Advances in Neural Information Processing Systems 28*, p. 2773–2781. Curran Associates, Inc.

WANG S., J.PAUL M. & DREDZE M. (2014). Exploring health topics in chinese social media : an analysis of sina weibo. *AAAI Workshop on the World Wide Web and Public Health Intelligence*, **23**, 20–23.

WANG Y., AGICHTEIN E. & BENZI M. (2012). Tm-lda : Efficient online modeling of latent topic transitions in social media. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, p. 123–131, New York, NY, USA : ACM.

WU Y., LIU S., YAN K., LIU M. & WU F. (2014). OpinionFlow : Visual Analysis of Opinion Diffusion on Social Media. *IEEE Transactions on Visualization and Computer Graphics*, **20**(12), 1763–1772.

ZENG Q. T. & TSE T. (2006). Viewpoint paper : Exploring and developing consumer health vocabularies. *JAMIA*, **13**(1), 24–29.

ZHAN Y., LIU R., LI Q., LEISCHOW S. & ZENG D. (2017). Identifying topics for e-cigarette user-generated contents : a case study from multiple social media platforms. *J Med Internet Res*, **19**(1), e24.

ZHU M., ZHANG Y., CHEN W., ZHANG M. & ZHU J. (2013). Fast and accurate shift-reduce constituent parsing.

# Etude d'une approche de Retour d'Expérience pour la découverte d'enseignements génériques dans le domaine humanitaire

Cécile L'Héritier[1], Sébastien Harispe[1], Abdelhak Imoussaten[1], Gilles Dusserre[1], Benoît Roig[2]

[1] LGI2P, IMT Mines Ales, Univ Montpellier, Ales, France
{prénom.nom}@mines-ales.fr
[2] EA7352 CHROME, Université de Nîmes, France
benoit.roig@unimes.fr

**Résumé** : Un intérêt croissant est exprimé par les organisations pour le développement d'approches visant à valoriser et tirer parti des expériences passées afin d'améliorer leurs processus de décision. Dans ce cadre de Retour d'Expérience, l'étude d'approches semi-automatisées pour la capitalisation et l'exploitation des connaissances revêt un intérêt central. Ce papier présente l'étude d'une approche au carrefour de plusieurs domaines : l'Ingénierie des Connaissances, l'Analyse Multicritère et l'Extraction d'Information. Elle repose sur une forme spécifique de raisonnement à partir de cas qui vise à établir une relation entre l'évaluation *a posteriori* de la performance de cas similaires et les caractéristiques des cas qui l'impactent. *In fine*, l'exploitation de ce lien a pour objectif d'identifier des enseignements génériques permettant de guider les processus décisionnels. Cette approche vise un cadre applicatif concret : la réponse logistique déployée par une ONG pour supporter des missions en situation d'urgence humanitaire.

**Mots-clés** : Retour d'Expérience, Ingénierie des Connaissances, Analyse Multicritère, ONG

## 1 Introduction

Le processus de Retour d'Expérience (RetEx) est assimilable en de nombreux points à une approche de gestion des connaissances qui vise à capitaliser, valoriser et diffuser tant que possible les connaissances explicites et tacites liées aux activités d'une organisation. L'objectif de cet exercice de gestion des connaissances est de garantir et d'améliorer le fonctionnement de l'organisation en tirant avantage des erreurs et succès passés (Liebowitz, 1999). Dans ce contexte, l'un des enjeux principaux des processus de RetEx est de développer et d'améliorer les techniques et outils permettant de tirer parti et de valoriser la connaissance d'experts du domaine. Cette connaissance est en effet souvent critique pour une organisation ; elle correspond au véritable substrat -connaissance forgée par l'expérience- à partir duquel les experts prendront leurs décisions. La mise en œuvre de démarches de RetEx est donc primordiale pour lutter contre la perte d'expertise et de connaissances au sein de l'organisation, et ainsi assurer la qualité et la performance des processus de l'organisation dans le temps. Cela est tout particulièrement le cas pour les processus impliquant des décisions qui nécessitent d'être étayées par une expertise, et argumentées par des éléments de connaissances identifiables. La nécessité et l'intérêt d'une telle démarche de RetEx est d'autant plus manifeste pour des organisations intervenant dans le cadre d'urgences humanitaires. La réponse à chaque situation d'urgence repose en effet avant tout sur de nombreux processus décisionnels -définition des moyens humains, compétences à mobiliser,

capacités matérielles à déployer... Ce sont ces choix délicats qui définiront la stratégie de réponse et conditionneront en grande partie le futur succès ou échec d'une mission. Du fait du contexte d'urgence, ces choix critiques doivent être rapidement formulés et dans la mesure du possible justifiés, et ce malgré une situation parfois insuffisamment définie et caractérisable. Dans la pratique, ces choix complexes reposent le plus souvent sur l'expertise et le savoir d'un nombre réduit d'experts. Ce patrimoine immatériel constitue alors le véritable capital de l'organisation, qu'il faut s'attacher à préserver et diffuser au sein de celle-ci. Cependant, cette connaissance est généralement difficile à exprimer et à formaliser (connaissance instinctive, *gut feeling*), ce qui rend la tâche de recueil de la connaissance particulièrement délicate.

Dans un contexte d'étude des processus de RetEx nous introduisons nos travaux préliminaires sur la définition d'une approche semi-automatisée permettant d'inférer des enseignements généraux *via* l'analyse d'expériences passées. Cette approche se situe au carrefour de plusieurs domaines : l'Ingénierie des Connaissances (IC), l'Analyse Multicritère et l'Extraction d'Information. De manière générale, l'approche ambitionne la définition de collaborations Homme-Machine pour la mise en place du RetEx. Elle repose sur une forme de raisonnement à partir de cas, et vise en particulier à établir une relation entre l'évaluation *a posteriori* de la performance de cas (missions ici) similaires et les caractéristiques des cas qui l'impactent. *In fine*, l'exploitation de ce lien a pour objectif d'identifier des enseignements généraux semblables à la connaissance qui gouverne certaines décisions expertes. L'étude de cette approche s'appuie sur un cas applicatif concret : la réponse logistique déployée par une ONG pour supporter des missions de distribution de nourriture, médicaments, abris et biens de première nécessité.

Le reste du papier est structuré comme suit. La section suivante présente des éléments d'information sur l'état de l'art relatif au RetEx ; nous nous intéresserons en particulier aux modes de représentation des expériences. La section 3 détaille les grandes étapes de l'approche proposée en soulignant son originalité. Une dernière section propose une discussion sur notre proposition, avant de conclure et d'évoquer les perspectives visées.

## 2 Etat de l'art et positionnement

### 2.1 Approches du Retour d'Expérience

La démarche de gestion des connaissances au sein d'une organisation fait référence à la manière dont celle-ci collecte, gère, et réutilise la connaissance qu'elle génère et acquiert, et ce en vue d'améliorer la performance et la qualité de ses processus. Cette démarche repose sur deux phases clés : la capitalisation -*capturer et stocker*- et l'exploitation de la connaissance permettant sa diffusion (Liebowitz, 1999). La démarche ou processus de RetEx, approche spécifique de gestion des connaissances, vise à analyser un événement passé dans l'objectif de réutiliser la connaissance qui en résulte et d'en tirer des enseignements pouvant avoir un impact positif sur les résultats de l'organisation e.g. éviter la reproduction d'erreurs, favoriser la diffusion de bonnes pratiques, etc., (Weber et al., 2001). Le vif intérêt des organisations pour la mise en œuvre de telles démarches, et la difficulté, pour des experts, de généraliser et de décontextualiser leurs connaissances (Kolb, 2000), se sont traduits par l'émergence d'un certain nombre d'approches, axées notamment sur la phase de capitalisation.

Parmi les approches les plus significatives, (Dieng-Kuntz et al., 2001) proposent de confronter le mode de représentation choisi et le type de mémoire produite. Ils distinguent ainsi des mémoires d'expériences et de mise en règles de l'expérience qui se font au travers de formulaires structurés constituant une base de données accessible par requêtes, e.g. les méthodes REX (Malevache et al., 1993) et MEREX (Corbel, 1997). D'autres approches, fondées sur des modèles de connaissances, tirent parti des techniques d'IC pour la constitution de mémoires d'activité. De telles mémoires cherchent à capturer la connaissance utilisée par un opérateur lors de la réalisation d'une tâche spécifique, et traduisent cette expertise au travers d'une série de modèles abstraits, e.g. CommonKads (De Hoog et al. 1996) et MKSM

(Ermine et al., 1996). L'inconvénient de ces modèles est parfois de présenter un niveau d'abstraction relativement élevé qui rend complexe leur mise en œuvre et leur adoption, ce qui nécessite souvent d'impliquer un intermédiaire -spécialiste en IC. Enfin, les approches de type référentiels métier visent à constituer un ensemble de glossaires, règles, et manuels opératoires, basés sur six types de connaissances : connaissances singulières, terminologiques, structurelles, comportementales, stratégiques et opératoires, e.g. CYGMA (Bourne, 1997). Ces référentiels métier, dans le cas des métiers de conception, ont permis la production de bases de connaissances exploitables *via* des algorithmes de raisonnement déductif dont le but est d'assister les activités de conception (Dieng-Kuntz et al., 2001). Ces méthodes de capitalisation reposent initialement sur un recueil d'informations issues d'entretiens avec des experts et sur l'analyse de documents. Elles sont par conséquent particulièrement coûteuses en termes de ressources humaines et chronophages (Ermine et al., 1996). Les approches orientées résolution de problèmes, plus complètes, peuvent être considérées dans le cadre de démarches de RetEx -en particulier le Raisonnement à Partir de Cas (RàPC), (Kolodner, 1993). Cette technique, ayant ses origines dans le raisonnement par analogie, permet d'adapter la solution d'un problème déjà résolu et renseigné dans une base de cas pour traiter une nouvelle situation. (Aamodt and Plaza 1994) définissent le cycle de RàPC par quatre étapes : la recherche, l'adaptation, la révision et l'apprentissage (certains auteurs considèrent aussi une phase d'élaboration visant à enrichir les données de la base). Dans la phase de recherche, le nouveau cas à résoudre est comparé aux cas stockés au moyen de mesures de similarités. Cette technique permet d'adapter des fragments de connaissances, mais ne permet pas de généraliser les connaissances spécifiques stockées dans la base. De plus, elle nécessite de disposer d'un nombre significatif de cas permettant de faire des analogies.

Dans le cadre du RetEx, de nombreux challenges sont donc ouverts pour automatiser l'extraction de connaissances à partir d'expériences passées, en particulier dans le domaine des ONG. Ce domaine soulève en effet des contraintes supplémentaires, du fait notamment de la difficulté d'obtenir des données pertinentes et précises, dans un monde où le témoignage oral prévaut. Par conséquent, dans ce contexte, prétendre à des analyses automatisées est un défi d'autant plus grand. De plus, le recueil et l'analyse de données peuvent rapidement être coûteux en termes d'investissement pour les experts, d'où la nécessité de minimiser l'information qui leur est demandée, et limiter ainsi l'effort cognitif et le temps investi.

## 2.2 Représentation de l'expérience

Une expérience peut être définie à partir de différents éléments d'information caractérisant le contexte dans lequel un événement intervient, l'analyse qui en est faite par les experts et la solution apportée au problème (Kamsu-Foguem et al., 2008). Le choix d'une représentation et d'un formalisme doit servir à valoriser l'expérience et permettre son exploitation, mais également faciliter l'identification des éléments de connaissance contenus dans chaque expérience. Dans cette optique, de nombreux formalismes de représentation des connaissances ont été étudiés.

Dans sa forme la plus simple, utilisée notamment dans les premières approches de RàPC, la formalisation des cas passés s'appuyait sur une représentation vectorielle (attribut-valeur) décrivant un problème et la solution associée (Bergmann et al., 2006). Bien que cette représentation permette d'évaluer facilement la similarité (proximité) des cas, la connaissance du domaine n'étant pas définie, elle ne permet pas de considérer la similarité sémantique des cas. Des approches textuelles sont également utilisées pour le RàPC ; elles tirent parti des techniques de Recherche d'Information, les mesures de similarité entre cas s'appuient sur l'occurrence de mots dans les documents qui les caractérisent -*keyword matching*-, (Bergmann, 2006). Le modèle des Frames (Minsky, 1975) a également été investigué dans le cadre du RàPC mais son manque de formalisme a pu être souligné comme une des faiblesses, notamment pour l'inférence de connaissances (Beler, 2008). Par extension, cet auteur s'est intéressé aux représentations orientées-objet relativement intuitives, car plus proches des représentations utilisées par les experts ; celles-ci ont été couplées à des modèles de croyances

permettant de prendre en compte l'incertitude liée à l'expérience -une application à la prévention des risques a été proposée. Enfin, dans le cadre de plusieurs travaux (Potes Ruiz et al., 2014), (Kamsu-Foguem et al., 2008) des formalismes et représentations basés sur l'utilisation d'ontologies et de graphes conceptuels sont adoptés. Les auteurs soulignent le réel intérêt présenté par les graphes conceptuels dans le cadre du RetEx, du fait notamment de leur représentation graphique facilement compréhensible, et de la possibilité d'appliquer des traitements automatiques. Ils présentent également un compromis vis-à-vis de l'expressivité et de la complexité des raisonnements qui peuvent être appliqués.

## 3 Une approche de RetEx basée sur le couplage de l'Ingénierie des Connaissances, l'Analyse Multicritère, et l'Extraction d'Information

L'approche de RetEx étudiée dans nos travaux s'articule autour de trois grandes étapes illustrées dans le schéma général proposé en Figure 1. Indépendamment de son caractère générique, l'approche est ici présentée au travers de son application au domaine humanitaire. La notion de mission fait référence de façon générique à une solution, une stratégie mise en œuvre en réponse à un problème dans un contexte donné. Dans notre contexte applicatif spécifique, les missions correspondent à la description des différentes missions de distributions menées par l'ONG avec laquelle nous collaborons. **A-** La première étape de traitement est dédiée à la valorisation des informations issues de l'expérience *via* la définition d'une base de connaissances RDF. Celle-ci a pour objectif de formaliser la connaissance associée aux différentes missions passées, e.g. lieu, stratégie déployée, contexte géopolitique et sanitaire… Cette étape conduit ensuite à deux phases d'exploitation distinctes avec pour objectif d'analyser et de tirer parti des expériences passées en vue de formuler des enseignements génériques. **B-** La première de ces deux étapes questionne, à l'aide d'interactions avec des experts du domaine, le succès/échec des missions. Cette étape s'appuie principalement sur les méthodes d'Analyse Multicritère qui accompagnent : (i) l'évaluation de la performance des missions, et (ii) l'identification d'un sous-ensemble de critères d'évaluation contribuant majoritairement à établir cette performance. **C-** La dernière phase d'exploitation s'intéresse au couplage entre les résultats issus de l'évaluation (B) et les observations (caractéristiques communes) pouvant être amenées par l'analyse de missions similaires formalisées (A), et ce en vue de tirer des enseignements pertinents.

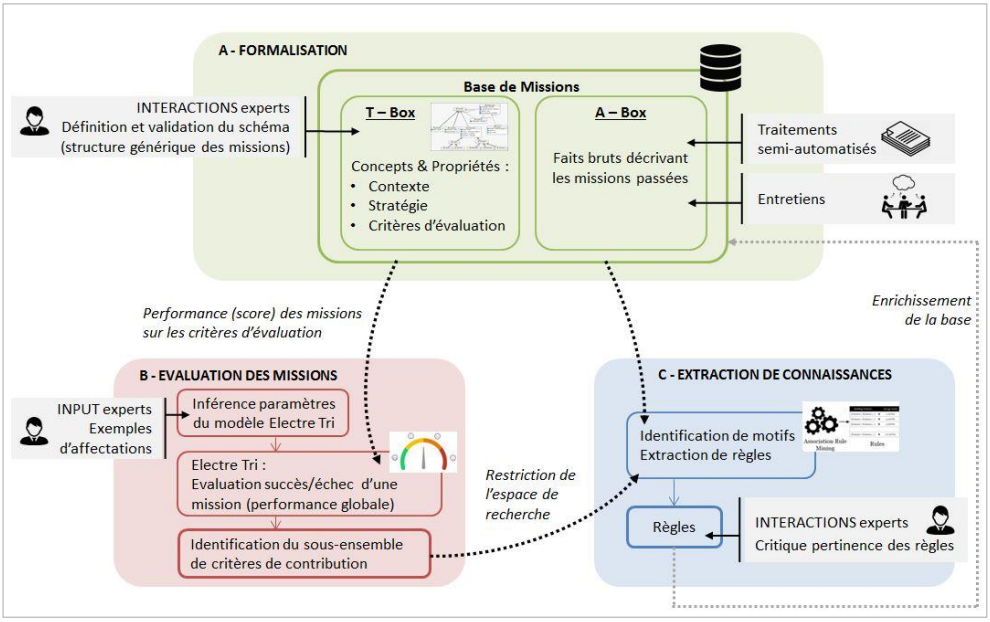



FIGURE 1- Schéma général de l'approche de RetEx étudiée illustrant les trois étapes principales.

Cette étape a donc pour but d'identifier des caractéristiques communes entre les missions pouvant expliquer la performance observée sur les critères d'évaluation, et en particulier ceux dont la contribution à la performance globale est significative. Des hypothèses d'enseignements pourront alors, dans l'idéal, être formulées à partir de la détection de ces règles. Des interactions fortes doivent être maintenues avec les experts et décideurs du domaine à chacune des étapes pour critiquer la pertinence des résultats. Des détails techniques sur les trois étapes introduites sont présentés ci-après.

### 3.1 Construction du référentiel de connaissance et formalisation des missions

L'objectif de cette étape est de définir la base de connaissances qui permettra de structurer les missions en s'appuyant sur une structure générique et standardisée -support nécessaire en vue de mener une analyse de RetEx. Une grande quantité de données, retraçant les décisions successives prises dans un contexte spécifique, peut être associée à chaque mission. C'est à partir de ces données que l'*expérience* doit être extraite. Celles-ci sont collectées auprès d'une multiplicité de sources (sources textuelles, bases de données, fichiers Excel, dires d'experts, retranscriptions) ; elles sont généralement fortement hétérogènes, très souvent exprimées en langage naturel, et par nature généralement imparfaites puisque entachées d'imprécisions, incomplètes, etc. Dans leur forme brute et faiblement structurée, les données sont alors difficilement exploitables ; le manque de formalisme rend leur analyse et exploitation *via* des traitements automatisés très difficile, voire inenvisageable. Par conséquent, la première étape de l'approche proposée a pour objectif de définir un cadre de formalisation permettant de valoriser tant que possible ces données. Cette valorisation est accompagnée par la constitution d'une base de missions exploitable, qui sera par la suite analysable par des techniques de traitements semi-automatisés qui permettront d'envisager la mise en œuvre d'analyses avancées. La base de missions est construite à l'aide de différents traitements (semi-) automatisés (e.g. extraction d'information, segmentation de textes, reconnaissance d'entités nommées) et éventuellement supervisés, e.g. modélisation de la connaissance exprimée en langage naturel par des experts. Les différents traitements servant la définition du référentiel de connaissances ne seront pas détaillés ci-après. Nous nous concentrerons plutôt sur certains aspects techniques associés à la formalisation de la notion de mission.

Différents formalismes étudiés dans le domaine de la Représentation des Connaissances permettent de structurer et de formaliser les informations issues de l'expérience en vue de les valoriser. Les avantages associés à une représentation formelle de la base des missions méritent d'être soulignés : mécanismes d'inférences permettant d'inférer de nouveaux faits qui viendront enrichir la base, mais aussi de vérifier sa consistance, possibilité de définir, au travers du modèle, de la connaissance *a priori* qui mérite d'être considérée pour mettre en correspondance des similitudes entre cas…

La représentation des missions choisie, actuellement en phase de test, s'appuie sur différents standards proposés par le W3C, e.g. spécifications associées au Web Sémantique : RDF(S), OWL et dans une moindre mesure, des règles SWRL. Le choix d'une représentation des connaissances basée sur OWL 2 est notamment supporté par le choix possible de profils spécifiques permettant des degrés de complexité/expressivité variables, et ainsi un compromis entre expressivité et complexité de raisonnement. Ces langages ont aussi l'avantage important d'être très largement étudiés et outillés (éditeurs, mise en correspondance facilitée avec des bases de données). L'utilisation de ces standards permet aussi une intégration plus simple avec des ontologies et bases de faits existantes, ce qui présente un réel avantage dans notre cas, notamment pour la caractérisation du contexte associé aux missions.

Les choix de modélisation effectués sont motivés par notre applicatif ; les concepts et propriétés définis permettent de retranscrire les missions passées, et en particulier les réponses logistiques mises en œuvre en vue d'en tirer des enseignements. La représentation des missions s'intéresse alors tout particulièrement aux éléments reflétant une stratégie et le processus décisionnel suivi sans s'affranchir du contexte d'intervention dont les choix

dépendent fortement. Par souci de simplification, la représentation proposée s'affranchit pour l'instant de l'aspect temporel (important par ailleurs). Le modèle proposé sera critiqué et validé au travers d'interactions avec les experts. Cette première étape conduit donc à une représentation formelle et objective des missions.

### 3.2 Cibler la connaissance pertinente

Cette section présente une approche basée sur l'Analyse Multicritère qui vise à réduire l'espace de recherche et, par conséquent, cibler la connaissance pertinente à rechercher lors de la mise en œuvre du RetEx. La démarche du RetEx a pour objectif d'extraire des enseignements génériques, en particulier au regard des choix qui ont été réalisés dans un contexte de mission donné. En d'autres termes, il s'agit, *a posteriori*, de solliciter les experts du domaine en vue de questionner la justesse des choix qui ont été faits, et de déterminer si ceux-ci se sont avérés bons ou mauvais au regard des objectifs visés. D'un point de vue méthodologique, cela implique d'évaluer la performance globale d'une mission et donc la stratégie mise en œuvre lors de sa réalisation. Etant donnée la multitude de critères à prendre en compte, nous utilisons pour cette évaluation une méthode issue de l'Analyse Multicritère, (Bouyssou & Roy, 1993). Une fois cette évaluation effectuée, il s'agira d'identifier, parmi les critères d'évaluation, quels sont ceux qui contribuent fortement à l'appréciation de succès/échec associée à une mission -une information difficile à obtenir *a priori*, pas toujours évidente à formuler par les experts ou fortement subjective et biaisée. Cette étape est centrale dans notre approche car elle vise la réduction de l'espace de recherche à partir duquel les enseignements génériques mériteront d'être extraits. Il est en effet important de souligner que l'espace de recherche, initialement de très grande taille, ne permet pas d'envisager l'identification d'enseignements génériques pertinents du fait du nombre réduit de missions capitalisables. Autrement dit, le faible nombre d'observations rend difficile l'identification de règles pertinentes sans cibler au préalable les critères d'intérêts. L'objectif sous-jacent est alors de déterminer les critères sur lesquels le processus de RetEx et l'analyse des missions doivent se focaliser en vue de tirer des enseignements pertinents.

Parmi les nombreuses méthodes d'Analyse Multicritère développées, nous avons opté pour la méthode de surclassement Electre Tri (Figueira et al., 2016). Cette méthode présente en effet des avantages majeurs pour notre contexte d'application (ONG et situations d'urgence), du fait notamment de : (i) sa capacité à prendre en compte l'imprécision intrinsèque aux jugements experts, (ii) son caractère non compensatoire, signifiant qu'une excellente performance sur un critère ne peut pas compenser la mauvaise performance sur un autre critère, (iii) l'existence de procédures d'inférence pour définir les paramètres du modèle Electre Tri, et ainsi minimiser la quantité d'information demandée aux décideurs et l'effort cognitif, (iv) la facilité de réutilisation du modèle défini.

Le principe d'Electre Tri est d'affecter une mission donnée dans une des catégories prédéfinies -chaque catégorie représente dans notre contexte un degré de réussite/performance spécifique, l'affectation obtenue est donc une forme d'évaluation. Dans une seconde phase, nous cherchons à analyser l'évaluation des missions qui a été faite afin d'identifier les facteurs expliquant le classement formulé (et ainsi réduire l'espace de recherche qui sera analysé par la suite). Cela revient à déterminer parmi les critères d'évaluation le sous-ensemble de critères contribuant le plus fortement au classement proposé et donc à la performance globale de la mission. Dans le cadre de nos travaux, nous avons d'ores et déjà proposé une approche pour l'identification de ce sous-ensemble de critères pertinents, détaillée dans (L'Héritier et al., 2018).

### 3.3 Extraction de règles pour l'identification d'enseignements génériques

Les deux premières étapes de traitement introduites ont respectivement amené la définition d'une base de connaissances formelles, et l'identification du sous-ensemble de critères d'intérêt pour l'identification de facteurs susceptibles d'influencer la performance d'une

mission. L'objectif de cette troisième phase est d'exploiter la base de connaissances de missions et les facteurs identifiés afin de formuler des enseignements génériques et d'intérêt pour le domaine humanitaire, i.e. de distinguer des liens de causalité entre des caractéristiques des missions, et les performances observées sur les critères impactant fortement l'évaluation de la performance globale des missions. L'hypothèse est faite ici d'un lien entre les performances globales des missions, liées aux valeurs de performance sur les critères, et des caractéristiques des missions qui impactent la valeur prise par ces critères. Afin d'identifier les caractéristiques communes entre les missions, nous proposons de nous appuyer sur des techniques d'analyse de données et d'extraction de règles permettant la découverte de *motifs* fréquents et de corrélations dans des bases de données/connaissances. Les règles ont l'avantage d'être facilement interprétables, et répondent parfaitement à l'objectif de RetEx visé par nos travaux (celles-ci peuvent notamment être exprimées facilement en langage naturel). Nous souhaitons ici de préférence nous concentrer sur les règles reliées aux critères identifiés au préalable comme étant d'intérêt. Nous nous concentrons en particulier sur l'extraction de règles exprimées sous la forme de clauses de Horn. Ces règles traduisent une implication de la forme $B \Rightarrow H$, où B est le corps de la règle formé d'une conjonction d'atomes, et H l'atome constituant la tête de la règle (conséquence de l'implication). Les atomes correspondent à des faits contenus dans la base de connaissance -faits qui expriment des éléments de connaissance sur les différentes missions-, ou à des abstractions de ces faits obtenues par la substitution des objets ou sujets par des variables. L'identification de ces règles est amenée par l'observation de motifs récurrents qui peuvent être généralisés -sous réserve que ces motifs vérifient des seuils permettant de critiquer leur pertinence (support et confiance notamment). L'extraction de règles dans des bases de connaissances comporte cependant certaines technicités. La base de connaissances considérée repose en effet sur l'hypothèse de monde ouvert par opposition à l'hypothèse de monde fermé adoptée dans les bases de données et classiquement considérée en analyse de données -un fait inconnu ne peut être considéré comme faux. L'hypothèse de monde ouvert est particulièrement importante dans le cadre de nos travaux (contexte d'informations incomplètes et de données manquantes). A ce jour, les règles sont extraites à l'aide de l'approche proposée par (Galárraga et al., 2015) ; celle-ci a été développée en tenant compte des difficultés induites par l'appréciation de l'hypothèse de monde ouvert en se basant sur l'hypothèse de complétude partielle tirant parti de la génération de contre-exemples. Les règles extraites sont contraintes, i.e. connectées et fermées (les variables des atomes apparaissent au moins deux fois dans la règle). Ces contraintes évitent notamment l'extraction de règles peu pertinentes (atomes sans rapport au sein de la règle, etc.) et permettent de restreindre d'autant plus l'espace de recherche des règles potentiellement pertinentes. L'objectif de cette étape est alors d'extraire les règles traduisant l'impact de conjonctions de propriétés (caractérisant les missions) sur les critères identifiés préalablement. Dans un contexte de collaboration Homme-Machine, des interactions fortes avec les experts sont envisagées afin de critiquer la pertinence des règles, de manière itérative, et ainsi orienter les hypothèses d'enseignements qui seront explorées.

## 4 Conclusion et perspectives

Dans le cadre du RetEx, nous étudions la définition d'une approche semi-automatisée permettant d'inférer des enseignements génériques *via* l'analyse d'expériences passées. Ces enseignements génériques aideront par la suite les processus décisionnels. L'approche proposée s'articule suivant trois étapes (1) la valorisation des informations issues de l'expérience *via* la définition d'une base de connaissances RDF. (2) L'évaluation du succès/échec d'une mission, et l'identification d'un sous-ensemble de critères qui contribue fortement à établir ce résultat. (3) Le couplage entre les résultats issus de l'évaluation et les caractéristiques communes des missions pouvant expliquer la performance.

Nous avons présenté ici l'aspect théorique de la démarche. Dans nos travaux futurs nous nous focaliserons sur les particularités des tâches de collecte puis de capitalisation des

informations issues de l'expérience. Ces tâches se confrontent d'une part, à la difficulté d'exprimer et de formaliser des connaissances liées à des expertises individuelles, souvent instinctives et/ou tacites ; et d'autre part à la multiplicité et l'hétérogénéité des sources d'informations ; informations, par nature, imprécises et incomplètes. Le contexte applicatif auquel l'approche envisagée est étroitement liée, soulève des contraintes et défis supplémentaires. En effet, le nombre de missions que nous serons amenés à analyser sera relativement réduit, ce faible nombre d'observations est impactant pour l'identification de motifs et l'extraction de règles -les métriques permettant d'évaluer la confiance/qualité des règles étant peu adaptées. Il faudra par conséquent étudier les alternatives possibles pour évaluer les règles et par extension les enseignements. L'interaction avec les experts du domaine afin de critiquer la pertinence des règles extraites est une première piste. Dans un contexte de collaboration Homme-Machine, des interactions avec les experts seront définies à chaque étape de l'approche : validation des modèles, critique de résultats, inférence du modèle de préférence du décideur *via* un ensemble d'apprentissage, etc.

## Références


AAMODT A. & PLAZA E. (1994). Case-Based Reasoning: Foundational Issues, Methodological Variations and System Approaches. *AICom- Artificial Intelligence Communications*, 7(1), 39–59.

BELER C. (2008). Modélisation Générique D'un Retour D'expérience Cognitif: Application À La Prévention Des Risques. *Thèse de doctorat, Institut National Polytechnique de Toulouse, France*.

BERGMANN R., KOLODNER J. & PLAZA E. (2006). Representation in Case-Based Reasoning. *The Knowledge Engineering Review*, 20(3), 209–13.

BOURNE C. (1997). Catégorisation et formalisation des connaissances industrielles. *Connaissances et savoirs faire en entreprise, Hermès*, 179–197.

BOUYSSOU D., ROY B. (1993). Aide multicritère à la décision : Méthodes et cas. *Economica*, Paris.

CORBEL J.C. (1997). Méthodologie de retour d'expérience: démarche MEREX de Renault, *Connaissances et Savoir-faire en entreprise, Hermès, Paris*, 93–110.

DE HOOG R., BENUS B., VOGLER M. & METSELAAR, C. (1996). The commonKADS Organization Model: Content, Usage and Computer Support. *Expert Syst. Appl*., 11(1), 29–40.

DIENG-KUNTZ R., CORBY O., GANDON F., GIBOIN A., GOLEBIOWSKA J., MATTA N., RIBIERE M. (2001). Méthodes et outils pour la gestion des connaissances, 2eme ed. Dunod.

ERMINE J., CHAILLOT M., BIGEON P., CHARRETON B., & MALAVIEILLE D. (1996). MKSM : Méthode pour la gestion des connaissances. *ISI, AFCET-Hermès*, vol. 4(4), 541–575.

FIGUEIRA J.R., MOUSSEAU V., ROY B. (2016). Electre methods. *In: Multiple Criteria Decision Analysis. Springer*, 155–185.

GALÁRRAGA L., TEFLIOUDI C., HOSE K., & SUCHANEK, F. M. (2015). Fast rule mining in ontological knowledge bases with AMIE+. *The VLDB Journal*, *24*(6), 707-730.

KAMSU-FOGUEM B., COUDERT T., BÉLER C., GENESTE L. (2008). Knowledge Formalization in Experience Feedback Processes: An Ontology-Based Approach. *Comp. Indust.* 59(7), 694-710.

KOLB D. (2000), Strategic Learning in a Knowledge Economy, The process of experiential learning, Butterworth-Heinemann, chap.15, 313–331.

KOLODNER J., (1993), Case Based-Reasoning, Morgan Kaufmann Publishers.

L'HÉRITIER C., IMOUSSATEN A., HARISPE S., DUSSERRE G., ROIG B. (2018). Identifying Criteria Most Influencing Strategy Performance: Application to Humanitarian Logistical Strategy Planning. *In Proceedings of the 17th International Conference, IPMU 2018.*

LIEBOWITZ J. (2001). Knowledge management and its link to artificial intelligence. *Expert Syst. Appl.*, 20(1), 1–6.

MALVACHE P. & PRIEUR P.(1993). Mastering corporate experience with the rex method. *In: Proceedings of ISMICK*, 93, 33-41.

MINSKY M. (1975). A framework for representing knowledge, *Psychol. Comp. Vis*. 211–277.

POTES RUIZ P., KAMSU-FOGUEM B. & GRABOT B. (2014). Generating Knowledge in Maintenance from Experience Feedback. *Knowledge-Based Systems 68*. Elsevier B.V, 4–20.

WEBER R., AHA D.W., BECERRA-FERNANDEZ I. (2001). Intelligent Lessons Learned Systems. *Expert Syst. Appl.*, 20(1), 17–34

# Exploration par apprentissage de discussions de personnes en détresse psychologique

Rémy Kessler[1], Nicolas Béchet[1], Gudrun Ledegen[2], Frédéric Pugnière-Saavedra[3]

[1] UNIV. BRETAGNE-SUD, UMR CNRS 6074, IRISA,
56017 Vannes, France
remy.kessler@univ-ubs.fr, nicolas.bechet@irisa.fr
[2] UNIV. RENNES II, PREFICS, EA 4246,
Campus Villejean
35043 Rennes, France
gudrun.ledegen@univ-rennes2.fr

[3] UNIV. BRETAGNE-SUD, PREFICS, EA 4246,
56017 Vannes, France
frederic.pugniere-saavedra@univ-ubs.fr

**Résumé** : Afin de s'adapter au mieux à la société, une association a développé une application de webchat permettant à toute personne d'exprimer et de partager ses préoccupations et ses malaises. Plusieurs milliers de conversations anonymes ont ainsi été réunies et forment un corpus inédit de récits sur la détresse humaine, les violences sociales. Nous présentons une méthode d'analyse de corpus combinant apprentissage non supervisé et *word embedding* afin de faire émerger les thématiques de cette collection particulière. Nous comparons la qualité de cette approche avec un algorithme standard de la littérature sur un corpus étiqueté et obtenons des résultats d'excellente qualité. Nous présentons une interprétation des regroupements obtenue sur cette collection particulière.

**Mots-clés** : word2vec, apprentissage non supervisé, word embedding

## 1 Introduction

Depuis les années quatre-vingt-dix, la souffrance sociale est une thématique qui fait l'objet d'une grande attention de la part de l'action publique et associative. Parmi les conséquences, figure l'explosion des lieux d'écoute ou des dispositifs sociotechniques de communication dont les finalités consistent notamment à modérer les diverses formes de souffrance par la libération de la parole dans un but thérapeutique Fassin (2004, 2006). Dans le cadre du projet METICS [1], une association de prévention du suicide a développé une application de *webchat* afin de répondre à ce besoin (Huet, 2015). Le *webchat* est un espace qui permet à toute personne d'exprimer et de partager avec un écoutant bénévole ses préoccupations et ses malaises. La principale spécificité de ce dispositif est son caractère non public et anonyme. Protégés par un pseudonyme, les écrivants sont invités à confier auprès d'un bénévole les aspects problématiques de leur existence. Plusieurs milliers de conversations anonymes ont ainsi été réunies et forment un corpus inédit de récits sur la détresse humaine. La finalité du projet METICS est de visibiliser les formes de souffrance ordinaires habituellement retranchées des espaces communs d'apparition et de saisir tant ses modalités d'énonciation que sa prise en charge au moyen des technologies numériques.

Dans le cadre de cette étude, nous souhaitons faire émerger de manière automatique les motifs de venue sur le chat des différents participants. En effet, même si l'association nous a fourni les thématiques abordées par l'ensemble des conversations (le travail, la solitude, la violence, le racisme, les addictions, les problèmes familiaux ou sentimentaux, etc.), le motif original de la venue sur le chat n'a pas été conservé. Dans la section suivante, nous présentons un état de l'art des différentes méthodes d'apprentissages comparables à notre proposition.

1. https://www.mshb.fr/projets_mshb/metics/2286/

La section 3 présente quelques statistiques sur les données tandis que la méthodologie est détaillée en section 4. La section 5 présente le protocole expérimental, une évaluation de notre système ainsi qu'une interprétation des regroupements finaux sur la collection de récits sur la détresse humaine.

## 2 Travaux connexes à notre proposition

La particularité de l'approche présentée dans ce papier est, d'un point de vue de l'utilisateur, de n'avoir à fournir que le libellé des classes à prédire. Ainsi, elle ne nécessite pas d'avoir un jeu de données étiquetées afin de prédire les différentes classes, c'est pourquoi elle est plus proche d'une méthode à base d'apprentissage non supervisé (ou semi-supervisé) que d'une méthode supervisée.

La littérature propose un certain nombre d'approches à base d'apprentissage non supervisé (ou clustering). L'idée du clustering est de regrouper des données non étiquetées dans un certain nombre de clusters, tel que des exemples similaires soient regroupés ensembles et ceux différents soient séparés. Pour une approche de clustering, le nombre de classes et la distribution des instances entre les classes ne sont pas connus *a priori* et le but est de trouver des regroupements significatifs. Les approches de clustering peuvent être classées selon le type de données fourni en entrée de l'algorithme et selon les critères de regroupement définissant la similarité ou la distance entre les données. Fraley & Raftery (1998) ont suggéré de diviser les algorithmes de clustering en deux catégories : les algorithmes hiérarchiques et ceux à base de partitionnement. Han & Kamber (2001) ont proposé de les catégoriser en trois catégories principales supplémentaires : les méthodes basées sur la densité, sur la modélisation et à base de grille.

L'algorithme de partitionnement des k-means fait partie des algorithmes de clustering les plus populaires, car il fournit un bon compromis entre la qualité de la solution obtenue et sa complexité de calcul (Arthur & Vassilvitskii (2007)). Même si k-means a été proposé pour la première fois il y a plus de 50 ans (MacQueen (1967)), il reste l'un des algorithmes les plus utilisés pour le clustering. En pratique, les k-means visent à trouver k centroïdes, un pour chaque cluster, minimisant la somme des distances de chaque instance de données par rapport à son centroïde respectif. Nous pouvons citer d'autres algorithmes à base de partitionnement comme les k-medoids ou PAM (Partition Around Medoids) qui est une évolution des k-means (Kaufman & Rousseeuw (1987)). Les approches hiérarchiques produisent quant à eux des clusters en partitionnant récursivement les données de manière descendante ou ascendante. Par exemple, dans une classification ascendante hiérarchique ou CAH (Lance & Williams (1967)), chaque exemple issu du jeu de données représente initialement un cluster. Ensuite, les clusters sont fusionnés, selon une mesure de similarité, jusqu'à ce que la structure arborescente souhaitée soit obtenue. Le résultat de cette méthode de clustering est appelé un dendrogramme.

Parmi les autres méthodes de clustering, les méthodes basées sur la densité supposent que les données appartenant à chaque cluster soient tirées d'une distribution de probabilité spécifique Banfield & Raftery (1993). L'idée est de faire croître un cluster donné tant que la densité dans le voisinage du cluster dépasse un certain seuil prédéfini. Les méthodes de classification basées sur un modèle reposent sur la découverte de descripteurs (ou caractéristiques) pour représenter chaque cluster. Les méthodes les plus utilisées pour ce type de méthodes sont les arbres de décision et les réseaux de neurones. Le plus populaire (qui sont à base de réseaux de neurones) sont les cartes de Kohonen ou self-organizing map - SOM (Kohonen (1982)). Finalement, les méthodes à base de grille partitionnent l'espace en un nombre fini de cellules qui forment une structure de grille.

Les approches à base d'apprentissage semi-supervisé tel que l'algorithme de propagation de libellés (Raghavan *et al.* (2007)) se rapprochent de la méthode proposée dans ce papier en se sens qu'elles consistent à utiliser un jeu de données d'apprentissage constitué de peu de données étiquetées et d'un nombre plus important de données non étiquetées afin de construire un modèle. Plus proche de la thématique de notre collection, Pestian *et al.* (2012) et Abboute *et al.* (2014) utilisent des approches supervisées pour détecter automatiquement

les personnes suicidaires dans les réseaux sociaux. Ils extraient des caractéristiques spécifiques pour entrainer différents classifieurs et compare les performances de leur système aux jugements de professionnels de la santé mentale. Plus récemment, une des tâches du challenge CLEF 2018 [2] était la détection des risques de dépression sur des textes écrits dans les médias sociaux Losada & Crestani (2016). Cependant, ces travaux et ce challenge impliquent des ensembles de données étiquetés, ce qui est la principale différence avec notre approche proposée (nous n'avons pas de jeu de données étiqueté).

## 3 Données et statistiques

L'association a fourni à l'équipe de recherche une collection de conversations entretenues entre les bénévoles et des appelants entre 2005 et 2015. La figure 1 présente un extrait anonymisé de conversation issue de cette collection.

```
...
Chat-association(21:35:55): Bonsoir
Appelant(21:34:14): Bsr, comment vivre avec un homme indecis?
Chat-association(20:32:33): c'est à dire?
Appelant(21:33:58): un homme qui veut divorcer un jour et un autre jour
qui est heureux d etre avec vous
Chat-association(20:34:35): Alors la question est peut-être que voulez-vous vous?
Appelant(21:35:54): je veux vivre heureu avec mon mari on vient de se marier
Appelant(21:37:03): donc je patiente j prends sur moi
Chat-association(20:38:12): lui comment il réagit?
Appelant(21:40:51): ms moi je suis son bouquet misere
Chat-association(20:41:12): "le bouquet misere"?
Appelant(21:42:17): bouc missaire
Chat-association(20:43:39): ok
Chat-association(20:44:59): vous avez le sentiment d'être un bouc émissaire
Appelant(21:55:40): oui
...
```

FIGURE *1 – Extrait d'une discussion issue de la collection METICS.*

Afin de réduire le bruit dans la collection, nous avons filtré l'ensemble des discussions contenant moins de 15 échanges entre un appelant et une personne de l'association, ces échanges étant généralement peu représentatifs (problème de connexion, demande d'information, etc. ). Nous observons des phénomènes linguistiques bien particuliers comme des émoticônes [3], des apocopes (par exemple « ado », « télé », « bi ») des acronymes, des fautes (orthographiques, typographiques, mots collés, et d'une très grande morphovariabilité et d'une créativité explosive (Kessler *et al.*, 2004)). Ces phénomènes doivent leur origine au mode de communication (direct ou semi-direct), à la rapidité de composition du message ou aux contraintes technologiques de saisie imposées par le matériel (terminal mobile, tablette, etc.).

En complément de cette collection, nous avons utilisé dans le cadre de ces travaux un sous-ensemble du corpus des textes du journal Le-Monde [4]. Ce sous-ensemble issu de la collection d'origine contient les articles filtrés en fonction de leurs étiquettes thématiques. Nous conservons ainsi les articles ayant pour thématique la télévision, la politique, l'art, la science ou encore l'économie. Le tableau 1 présente quelques statistiques descriptives de ces deux collections.

## 4 Méthodologie

### 4.1 Vue d'ensemble du système

La figure 2 présente une vue d'ensemble du système dont les étapes seront détaillées dans le reste de la section. Au cours d'une première étape (module ①), nous appliquons différents

2. http://early.irlab.org/
3. Symboles utilisés dans les messages pour exprimer les émotions, exemple le sourire :-) ou la tristesse :-(
4. http://www.islrn.org/resources/421-401-527-366-2/

| Collection | METICS | Le-Monde |
|---|---|---|
| Nombre total de documents | 17 594 | 205 661 |
| *avant prétraitements linguistiques* | | |
| Nombre total de mots | 12 276 973 | 87 122 002 |
| Nombre total de mots différents | 158 361 | 419 579 |
| Nombre moyen de mots par conversation/doc. | 698 | 424 |
| *après prétraitements linguistiques* | | |
| Nombre total de mots | 4 529 793 | 41 425 938 |
| Nombre total de mots différents | 120 684 | 419 006 |
| Nombre moyen de mots par conversation/doc. | 257 | 201 |

TABLE *1 – statistiques du corpus METICS et du corpus Le-Monde.*

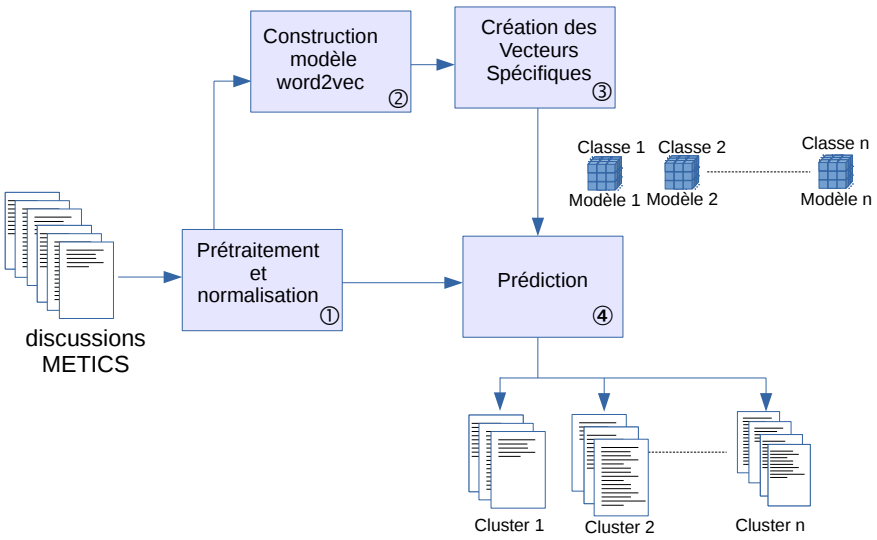



FIGURE *2 – Vue d'ensemble du système*

pré-traitements linguistiques à chacune des discussions. Le module suivant (②) constitue un modèle de *word embedding* à partir de ces discussions tandis que le troisième module (③) utilise ce modèle afin de créer des vecteurs spécifiques. Le dernier module ④ effectue une prédiction pour chaque discussion avant de séparer l'ensemble de la collection en regroupement en fonction de la classe prédite.

## 4.2 Pré-traitements et normalisation

Nous effectuons au préalable une extraction du contenu textuel de chaque discussion. Au cours de l'étape ①, une normalisation des textes est effectuée afin d'améliorer la qualité du processus. On supprime ainsi les accents, les caractères particuliers tels que « - », « / », « ( ) » . Différents processus linguistiques sont utilisés afin de réduire le bruit dans le modèle : les chiffres et nombres (numériques et/ou textuels), les symboles spéciaux ainsi que les termes contenus dans un anti-dictionnaire classique. Un processus de lemmatisation avait été intégré lors des premières expériences, mais il s'est avéré peu performant compte tenu de variations typographiques décrites en section 3. Au cours d'expériences préliminaires, nous avons tenté un filtrage permettant de prendre en compte uniquement l'un des rôles dans la conversation (appelant ou personne de l'association). Nous attribuons les faibles performances obtenues au rôle complémentaire de chaque intervenant (explicitation du message suivi d'une refor-

mulation) pour guider le système.

### 4.3 Construction d'un modèle word2vec

L'étape suivante de notre système consiste à construire un modèle de *word embedding* à l'aide de word2vec (Mikolov *et al.* (2013)). Cela consiste à projeter chaque mot de notre corpus dans un modèle vectoriel afin d'obtenir une représentation sémantique de ces derniers. Ainsi, des mots apparaissant dans des contextes similaires posséderont une représentation vectorielle relativement proche. Outre l'information sémantique, un des avantages d'une telle modélisation est le fait de produire des représentations vectorielles de mots, en fonction du contexte dans lesquels ils sont rencontrés. Ainsi, certains mots proches d'un terme *t* dans un modèle appris à partir d'un corpus *c1* pourront être très différents de ceux issus d'un modèle appris à partir d'un corpus *c2*. Nous observons par exemple sur la table 2 que les dix premiers mots proches du terme "ado" varient en fonction du corpus utilisé. Cet exemple montre également que l'utilisation d'un modèle générique type Wikipedia n'est pas pertinent dans notre cas, le corpus de l'association étant bruité et contenant un certain nombre d'apocopes, abréviations ou acronymes tels qu'"ado", "prob", ou encore "tv". Différents paramètres ont été testés et la configuration obtenant les meilleurs résultats a été conservée [5].

| corpus | mots |
|---|---|
| METICS | adolescente, jeune, adolescence, 15ans, lycéenne, gaté, adolescent, gamin, gâtée, 14ans |
| Le-Monde | kyrou, fun, fm, nrj, marmelade, ouie, tropic, skyrock, difool, garnement |

TABLE *2 – Dix mots les plus proches du terme "ado" selon le type de corpus en apprentissage*

### 4.4 Constitution des vecteurs spécifiques et prédiction

Au cours de cette étape, nous constituons des vecteurs contenant des termes choisis à l'aide du modèle word2vec construit au cours de l'étape 4.3. Pour chacune des thématiques contenues dans la collection, nous construisons ainsi un modèle linguistique spécifique en effectuant un plongement de mots (*word embedding*) afin de reconstruire le contexte linguistique de chacune des thématiques. Nous observons par exemple que les termes les plus proches de la thématique "travail" sont : « chômage », « boulot » , « boulo » , « stress » . De même, pour la thématique "addiction", nous observons les termes : « canabis », « alcoolisme » , « drogue » et « héroine » . Nous utilisons par la suite ce contexte afin de construire un vecteur, contenant la distance $dist_c(i)$ entre chaque terme $i$ et la thématique $c$. Chacun de ces modèles étant indépendant, un même terme peut ainsi apparaître dans plusieurs modèles. On observe ainsi que le mot "stress" présent dans le vecteur "suicide" et dans celui de "travail", cependant le poids associé est différent. Nous avons fait varier la taille de ces vecteurs entre 20 et 1000 et les meilleurs résultats ont été obtenus avec une taille de 400. Au cours de la dernière étape ④, le système calcule un score $S_c$ pour chaque discussion et pour chaque classe en fonction de chaque modèle linguistique tel que :

$$S_c(d) = \sum_{i=1}^{n} tf(i) \cdot dist_c(i) \tag{1}$$

avec $i$ le terme considéré, $tf(i)$ la fréquence de $i$ dans la collection, et $dist_c(i)$ est la distance entre le terme $i$ et la thématique $c$. On attribue au final la classe ayant obtenu le score le plus élevé.

5. Les meilleurs résultats ont été obtenus avec les valeurs de paramètres suivantes : taille des vecteurs : 700, taille de la fenêtre glissante : 5, fréquence minimale : 10, méthode de vectorisation : skip-gram, utilisation d'une fonction softmax hiérarchique pour l'apprentissage du modèle.

## 5 Expériences et résultats

### 5.1 Protocole expérimental

Afin d'évaluer la qualité des clusters obtenus, nous avons utilisé un sous-ensemble des textes du journal Le-Monde, décrit en section 3, chaque article possédant une étiquette en fonction de la thématique. Dans le cadre de ces expériences, nous avons configuré l'approche des vecteurs spécifiques (VS) avec les paramètres optimaux, tels que définit en section 4.3 et 4.4. Afin de tester l'influence particulière de ce paramètre, nous avons également testé les vecteurs spécifiques sans la pondération. Afin de montrer la difficulté de la tâche, nous comparons notre système avec une *baseline* sous forme de tirage aléatoire, ainsi qu'avec l'algorithme des k-means (MacQueen (1967)), couramment utilisé dans la littérature tel que mentionné en section 2. Afin d'alimenter l'algorithme des k-means, nous avons transformé notre collection initiale en une matrice de type *bag of words* (Manning & Schütze (1999)) ou chaque conversation est décrire par la fréquence des mots qui la compose. Chacune des expériences a été évaluée en utilisant les mesures classiques de Précision, Rappel et F-score des documents bien classés, moyennés sur toutes les classes (avec $\beta = 1$ afin de ne privilégier ni la précision ni le rappel (Goutte & Gaussier (2005))). L'algorithme des k-means n'associant pas d'étiquette au regroupement produit, nous avons calculé de manière exhaustive l'ensemble des solutions pour ne garder que celle obtenant le F-score le plus élevé.

### 5.2 Résultats

| | Precision | Rappel | F-score |
|---|---|---|---|
| Sans prétraitements linguistiques | | | |
| Baseline | 0.18 | 0.16 | 0.17 |
| k-means | 0.23 | 0.20 | 0.22 |
| VS sans pondération | 0.54 | 0.50 | 0.52 |
| Vecteurs Spécifiques (VS) | 0.53 | 0.54 | 0.53 |
| Avec prétraitements linguistiques | | | |
| k-means | 0.30 | 0.21 | 0.25 |
| VS sans pondération | 0.55 | 0.51 | 0.53 |
| Vecteurs Spécifiques(VS) | **0.54** | **0.54** | **0.54** |

TABLE *3 – Ensemble des résultats obtenus par chaque système.*

Le tableau 3 présente une synthèse des résultats obtenus avec chaque système. On observe dans un premier temps que la baseline obtient un score très faible, mais qui reste relativement proche de l'aléatoire théorique (0,2) compte tenu du nombre de classes. L'utilisation des prétraitements linguistiques apporte peu individuellement, mais permet d'améliorer globalement les résultats des autres expériences. L'algorithme des k-means obtient des résultats légèrement meilleurs en termes de F-score, mais reste faible. Les vecteurs spécifiques obtiennent d'excellents résultats qui surpassent les autres systèmes avec un F-score de 0,54. L'exécution sans pondération montre que celle-ci permet d'améliorer légèrement le rappel.

### 5.3 Analyse des clusters

L'objectif initial de ses travaux étant l'exploration de la collection METICS, nous appliquons l'ensemble du processus avec l'approche des vecteurs spécifiques afin de catégoriser automatiquement l'ensemble des conversations du corpus. Dès lors, l'interprétation des clusters obtenus est réalisée avec l'Allocation de Dirichlet latente (ou latent Dirichlet allocation - LDA, Hoffman *et al.* (2010)) afin d'obtenir le sujet dominant de chaque cluster. Nous avons par la suite associé les poids à chacun des termes en fonction de chaque cluster et regroupé les mots-clés thématiques les plus significatifs dans le tableau 4. Ce dernier croise

les clusters avec deux types de matrices de discours : celles qui annoncent des directions sémantiques significatives en termes de présence (classées du rouge [très significatif] au bleu [ moins significatif]) et celles qui reformulent globalement des éléments individuels. La méthode d'analyse utilisée semble opératoire puisque d'une part, sur 17 étiquettes de clusters préétablis, 10 comportent des désignations très pertinentes, 4 (*psy, adolescence, alcool*) le sont moins et seulement 3 (*handicap, travail, racisme*) ne le sont pas du tout. Le fait que ces trois dernières thématiques ne soient pas pertinentes, bien qu'ils soient des vecteurs de mal-être bien identifiés, est à mettre en lien avec le public particulier qui pratique le chat : pour environ 3/4 constitué de jeunes filles, le "travail" ne les concerne pas encore, et des thématiques "handicap" et "racisme" sont devancées par d'autres comme "solitude", "violence", "viol". Cette méthode d'analyse permet d'autre part de faire ressortir des univers sémantiques et des regroupements thématiques pour mettre en mots le mal-être et pour expliquer pourquoi il y a du mal-être chez le scripteur. L'entrée dans ce corpus par les clusters permet notamment de mettre au jour des embrayeurs fonctionnant comme des routines discursives significatives (Née *et al.*, 2014) qui annoncent et qui reformulent ce qui ne va pas.

| | Comment dire "ce qui ne va pas" | | | | | |
|---|---|---|---|---|---|---|
| | du côté de l'annonce | | | du côté d'une forme de reformulation abstraite | | |
| **cluster** | peur | psy | confiance | chose | difficile | problème |
| maladie | 1,78 | 1,71 | 1,56 | | 1,54 | |
| adolescence | 1,71 | 1,57 | 1,61 | 1,46 | 1,47 | |
| solitude | 1,69 | 1,64 | 1,58 | 1,52 | 1,55 | 0,22 |
| suicide | 1,67 | 1,71 | 1,54 | 1,51 | | |
| rupture | 1,66 | 1,56 | 1,55 | 1,5 | 1,52 | |
| violence | 1,62 | 1,57 | 1,49 | 0,41 | 1,43 | |
| travail | 1,61 | 1,63 | 1,57 | 1,47 | 1,46 | |
| viol | 1,59 | 1,7 | 1,44 | 1,42 | 1,4 | |
| angoisse | 1,56 | 1,5 | 1,43 | 1,35 | 1,36 | |
| famille | 1,54 | 1,5 | 1,47 | 0,39 | 1 | |
| relation | 1,08 | 1 | 1,01 | 0,94 | 0,91 | |
| alcool | 0,89 | 0,88 | 0,27 | 0,79 | | 0,5 |
| deuil | 0,88 | 0,96 | | 0,77 | 0,77 | |
| racisme | 0,66 | 0,5 | | 0,63 | | |

TABLE *4 – Répartition des routines discursives par cluster.*

Dans la table 4, la peur, le psy et la confiance sont des désignations présentes pour chaque cluster avec un rang largement significatif ; pour autant le scripteur exprime-t-il toujours la peur quand il écrit, « j'ai peur d'être malade » ? Ces désignations ne participent-elles pas à ouvrir et à construire des sphères de significations autour de ces mots pivots ? Inversement, avec un rang inférieur, mais également significatif, les désignations chose, difficile, problème sont plus vagues, mais plus reformulantes pour reprendre les éléments qui participent à écrire ce qui ne va pas.

## 6 Conclusion et travaux futurs

Nous avons présenté dans cet article une approche non supervisée permettant d'explorer une collection de récits sur la détresse humaine. Cette approche utilise un modèle de *word embedding* afin de construire des vecteurs contenant uniquement du vocabulaire issu du contexte linguistique du modèle. Nous avons évalué la qualité de l'approche sur une collection étiquetée avec des mesures classiques. L'analyse détaillée a montré des résultats de très bonnes qualités (Fscore moyen de 0,54), comparativement aux autres systèmes testés. Cette méthode d'analyse a permis d'autre part de faire ressortir des univers sémantiques et des regroupements thématiques.

Nous envisageons dans un premier temps d'étudier plus en détail l'influence de chacun des paramètres sur les résultats obtenus. Nous envisageons par ailleurs afin de pouvoir attribuer plusieurs étiquettes à chaque discussion, ce qui permettrait de prendre en compte les chevauchements thématiques. L'analyse conforte l'approche par cluster pour faire ressortir les

traits définitoires de ce type de production de discours et pour en révéler un fonctionnement interne. Cette entrée par les routines discursives n'est qu'un exemple qui permettra ensuite d'aborder d'autres explorations avec notamment une focale sur les formes argumentatives et sur les formes d'intensité.

## Références


ABBOUTE A., BOUDJERIOU Y., ENTRINGER G., AZÉ J., BRINGAY S. & PONCELET P. (2014). Mining twitter for suicide prevention. In *Natural Language Processing and Information Systems : 19th International Conference on Applications of Natural Language to Information Systems, NLDB 2014, Montpellier, France, June 18-20, 2014. Proceedings*, p. 250–253 : Springer International Publishing.

ARTHUR D. & VASSILVITSKII S. (2007). K-means++ : The advantages of careful seeding. *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, p. 1027–1035.

BANFIELD J. D. & RAFTERY A. E. (1993). Model-based gaussian and non-gaussian clustering. In *Biometrics*, volume 49, p. 803–821.

FASSIN D. (2004). Et la souffrance devint sociale. In *Critique, 680(1)*, p. 16–29.

FASSIN D. (2006). Souffrir par le social, gouverner par l'écoute. In *Politix, 73(1)*, p. 137–157.

FRALEY C. & RAFTERY A. E. (1998). How many clusters ? which clustering method ? answers via model-based cluster analysis. *The Computer Journal*, p. 578–588.

GOUTTE C. & GAUSSIER É. (2005). A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. *ECIR 2005*, p. 345–359.

HAN J. & KAMBER M. (2001). Data mining : Concepts and techniques. *Kaufmann Publishers, USA*.

HOFFMAN M., BACH F. R. & BLEI D. M. (2010). Online learning for latent dirichlet allocation. In J. D. LAFFERTY, C. K. I. WILLIAMS, J. SHAWE-TAYLOR, R. S. ZEMEL & A. CULOTTA, Eds., *Advances in Neural Information Processing Systems 23*, p. 856–864.

HUET R. (2015). La voix des naufragés. Dire sa souffrance dans des associations d'écoute et de prévention du suicide. *Communication et langages. 2015/4 (186).*

KAUFMAN L. & ROUSSEEUW P. (1987). *Clustering by Means of Medoids*. Delft University of Technology : reports of the Faculty of Technical Mathematics and Informatics.

KESSLER R., TORRES J.-M. & EL-BÈZE M. (2004). Classification thématique de courriel par des méthodes hybrides. *Journée ATALA sur les nouvelles formes de communication écrite*.

KOHONEN T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, (1), 59–69.

LANCE G. N. & WILLIAMS W. T. (1967). A general theory of classificatory sorting strategies1. hierarchical systems. *The Computer Journal*, (4), 373–380.

LOSADA D. & CRESTANI F. (2016). A test collection for research on depression and language use. In *Proc. of Experimental IR Meets Multilinguality, Multimodality, and Interaction, 7th International Conference of the CLEF Association, CLEF 2016*, p. 28–39, Evora, Portugal.

MACQUEEN J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1 : Statistics*, p. 281–297, Berkeley, Calif. : University of California Press.

MANNING C. D. & SCHÜTZE H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA : MIT Press.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. & DEAN J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS'13*, p. 3111–3119, USA : Curran Associates Inc.

NÉE E., SITRI F. & VENIARD M. (2014). Pour une approche des routines discursives dans les écrits professionnels. *Congrès Mondial de Linguistique Française, DOI 10.1051/shsconf/20140801195.*

PESTIAN J. P., MATYKIEWICZ P., LINN-GUST M., SOUTH B., UZUNER O., WIEBE J., COHEN K. B., HURDLE J. & BREW C. (2012). Sentiment analysis of suicide notes : A shared task. *Biomedical Informatics Insights*, **5s1**, BII.S9042.

RAGHAVAN U. N., ALBERT R. & KUMARA S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical review. E, Statistical, nonlinear, and soft matter physics*, p. 036106.

# Différentes méthodologies pour la conception d'ontologies

Longtemps une limite évoquée quant à l'utilisation des ontologies dans des applications dédiées était le temps de conception des ontologies et la haute expertise nécessaire pour assurer leur qualité et donc l'efficacité des dites applications. C'est pourquoi de nombreux travaux de ces dernières années s'intéressent à la construction d'ontologies. Celle-ci peut être semi-automatique comme le propose le premier article qui ambitionne de créer un modèle de connaissance à partir de tables statistiques. La construction peut se faire suivant une méthodologie particulière comme le propose les autres articles pour la création d'ontologies modulaires, la construction d'une ontologie des objets de la connaissance scientifique ou la création d'une ontologie du domaine du *sourcing*.

# D'un modèle statistique à un modèle de connaissance : retour d'expérience

Rabia Azzi[1], Sylvie Despres[1], Jérôme Nobecourt[1]

UNIVERSITÉ PARIS 13, SORBONNE PARIS CITÉ, LIMICS,(U1142), INSERM, Sorbonne Universités, UPMC Université Paris 6, 74 rue Marcel Cachin F-93017 Bobigny cedex, France
prenom.nom@univ-paris13.fr

**Résumé** :
Les modèles statistiques sont couramment représentés sous forme textuelle, tabulaire et graphique dans des documents (rapports, articles, affiches et présentations) qui sont le plus souvent en format PDF. Même si ce format rend l'accès à l'information plus difficile, il est intéressant de traiter directement le fichier PDF. Dans cet article nous proposons une approche permettant le passage d'un modèle statistique de connaissances à un modèle de connaissances qui soit visualisable afin d'en permettre une exploitation plus aisée. Notre approche consiste à : (i) extraire les informations pertinentes sous forme de triplets RDF ; (ii) organiser les triplets pour construire un modèle conceptuel ; (iii) visualiser dynamiquement le modèle obtenu. Nous nous focalisons sur les deux premières étapes de la méthodologie.

**Mots-clés** : Modèle statistique, Tableau statistique, Modèle conceptuel, Extraction sémantique d'information, RDF

## 1 Introduction

Dans cet article, nous traitons de l'automatisation de l'interprétation de résultats d'approches statistiques publiées au format PDF. Les questions soulevées sont : (1) comment extraire l'information et la sémantique existante dans ces documents ; (2) comment traduire les informations extraites vers un modèle conceptuel ; (3) comment exploiter le modèle obtenu pour le rendre dynamique et interactif.

La démarche statistique, indispensable pour extraire des connaissances à partir de données, est utilisée dans presque tous les domaines de l'activité humaine : ingénierie, management, économie, biologie, informatique, etc. (Saporta, 2011). Les données sont collectées, exploitées, analysées et enfin souvent présentées sous formes tabulaires et graphiques afin de faciliter leur interprétation.

Un tableau statistique est un ensemble de cellules organisées en lignes et colonnes, contenant des données chiffrées. Un tel tableau peut être : (1) à une entrée, il permet l'étude d'un caractère d'une population ; (2) à double entrée, il permet d'étudier simultanément deux caractères d'une population. Quelle que soit la structure du tableau, l'approche de lecture des informations qu'il contient est identique. Il faut dans tous les cas étudier le titre, la source et comprendre les intitulés des lignes et des colonnes. Les informations représentées dans les tableaux sont fortement connectées, en particulier les liens entre les éléments du tableau sont implicites et nécessitent une explicitation. Cependant, comme toutes les données sont d'égale importance, il n'est pas évident d'identifier et de sélectionner les informations essentielles (Junyong & Sangseok, 2017). Notre objectif est de proposer un traitement semi-automatique pour l'interprétation de ces tableaux. En effet, ils constituent un des moyens les plus couramment utilisés pour présenter et structurer les informations. Ce traitement nécessite d'extraire les données qu'il contiennent et de les représenter dans un modèle facilitant leur compréhension. En outre, la plupart des documents sont publiés au format PDF à partir duquel l'extraction d'information est fastidieuse. De nombreuses approches ont été proposées pour extraire des informations à partir de documents PDF. Elles exploitent, soit le format des balises tel que HTML et XML, soit un format de texte brut pour l'extraction d'information (Riaz *et al.*, 2016), (Klampfl & Kern, 2015).

Dans cet article, nous décrivons une méthode permettant de traduire des tableaux statistiques au format PDF vers un modèle conceptuel guidant l'interprétation des informations

qu'il contient. Nous utilisons les techniques d'extraction d'information à partir de documents PDF et le modèle de triplets RDF [1] afin de structurer les informations extraites. Cette approche permet en effet d'explorer la manière de représenter des informations sans identifier initialement un vocabulaire source pour les prédicats (Powell, 2015). D'un point de vue applicatif, RDF utilise des identifiants uniques pour les ressources et chaque triplet correspond à la déclaration d'un fait.

Cet article est organisé comme suit : dans la section 2, nous introduisons la démarche de présentation des résultats en statistique. Dans la section 3, nous présentons l'approche générale et nous nous focalisons sur les deux premières étapes de la méthodologie (extraction des connaissances sous forme de triplets RDF et construction d'un modèle conceptuel). Dans la section 4 et 5, nous décrivons l'expérimentation, les résultats obtenus et leur évaluation. Dans la section 6, nous concluons et présentons des perspectives d'investigations complémentaires.

## 2 Présentation des résultats en statistique

Le mot statistique [2] désigne à la fois un ensemble de données d'observations et l'activité consistant en leur recueil, leur traitement et leur interprétation. Pour (Saporta, 2011), faire de la statistique suppose l'étude d'un objet ou d'un ensemble d'objets sur lesquels des caractéristiques appelées « variables » sont observées. La notion fondamentale en statistique est celle de population qui correspond à un groupe ou à un ensemble d'objets équivalents. Les objets sont appelés individu ou unité statistique. L'étude de tous les individus d'une population finie correspond à un recensement. Lorsqu'une partie de la population est observée, il s'agit d'un sondage, la partie étudiée étant désignée comme l'échantillon. Ainsi, chaque individu d'une population est décrit par un ensemble de caractéristiques appelées variables ou caractères. Ces variables peuvent être classées comme des variables quantitatives ou numériques (par exemple, taille, poids, etc.) ou comme des variables qualitatives s'exprimant par l'appartenance à une catégorie (par exemple, catégorie socio-professionnelle, etc.).

La démarche statistique comporte usuellement les étapes suivantes :

— le recueil qui consiste à collecter les données, selon deux grandes méthodologies : les sondages et les plans d'expériences ;
— l'exploration qui consiste à synthétiser, résumer, structurer l'information contenue dans les données ;
— l'inférence qui consiste à étendre les propriétés constatées sur l'échantillon de la population toute entière et à valider ou infirmer des hypothèses *a priori* ou formulées après une phase exploratoire ;
— la modélisation qui consiste généralement à rechercher une relation approximative entre une variable et plusieurs autres. Les modèles souvent utilisés sont la régression linéaire, le modèle linéaire général et la méthode de discrimination.

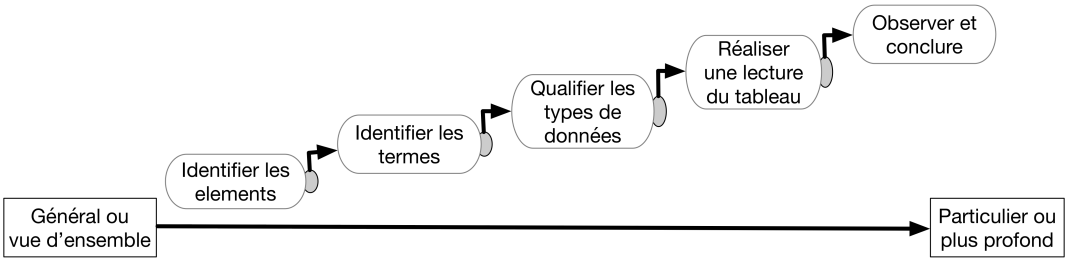



FIGURE *1 – Etapes d'analyse d'un tableau statistique*

1. https ://www.w3.org/RDF/
2. http ://www.larousse.fr/dictionnaires/francais/statistique/74516

Les méthodes visuelles sont largement utilisées en statistique pour présenter les résultats de manière claire et concise. Il existe plusieurs formes possibles de présentation visuelle des résultats : tableaux, graphiques, histogrammes, diagrammes, etc. Parmi ces formats, les tableaux sont les plus utilisés, car l'information est disposée de manière à mettre en évidence les relations entre les données.

Il existe trois types de tableaux : (i) les tableaux de données (les premiers construits) qui sont généralement « de grande taille » puisqu'ils comptent autant de lignes que de sujets étudiés ; (ii) les tableaux de distribution de variables (les plus connus) sont obtenus par regroupement des cases identiques figurant dans les colonnes et décrivent la distribution d'une variable ; (iii) les tableaux de contingence sont constitués par croisement de deux variables renseignées.

Les tableaux statistiques permettent d'organiser et de présenter les données ou les résultats en regroupant des informations de même nature. Cependant pour les exploiter, une démarche rigoureuse doit être suivie. Le grand principe d'analyse d'un tableau en statistique (voir Figure 1) est d'adopter une démarche allant du général au particulier. Cette démarche comporte les étapes suivantes :

— identifier les éléments : consiste à identifier le titre, la source de l'étude, la nature du tableau, etc. Chacun de ces éléments est porteur d'information (par exemple, le titre peut renseigner sur l'idée, la variable expliquée, etc.) ;
— identifier les termes : consiste à identifier les termes figurant dans le titre, les colonnes et les lignes du tableau, etc. ;
— qualifier les types de données : consiste à qualifier le type de données contenu dans le tableau en prenant en compte les unités (par exemple, des pourcentages, des probabilités, etc.) ;
— réaliser une lecture du tableau : consiste à appliquer deux règles communes de lecture à tous les tableaux. La première règle consiste à construire une paraphrase en débutant la lecture en se plaçant sur une ligne et en poursuivant par celle des colonnes utiles à l'analyse. La seconde consiste à répéter la première règle sur plusieurs lignes du tableau pour vérifier la pertinence des relations ;
— observer : consiste à tirer des conclusions à partir du tableau. Par exemple, identifier des relations entre variables (cause/effet), des valeurs extrêmes, des tendances, etc.

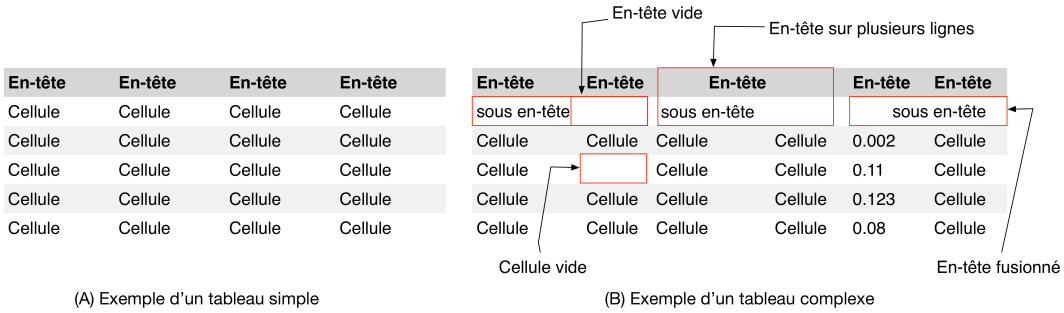



FIGURE *2 – Exemple de tableau simple et complexe*

Si une analyse visuelle permet à un être humain de reconnaître et comprendre facilement les tableaux, la situation est différente quant il s'agit d'un ordinateur. Un tableau est constitué de cellules d'en-tête et de cellules contenant des données. Cette structuration permet de définir la relation entre les cellules et fournit un contexte aux utilisateurs. Selon (Yeon-Seok & Kyong-Ho, 2008), il existe deux types de tableaux (voir Figure 2) : (1) les tableaux simples (voir Figure 2-A) comportant au maximum une ligne d'en-tête subdivisée en colonnes d'en-têtes. La colonne d'en-tête précise le type d'information qu'elle contient. Il n'y a pas de cellules fusionnées dans un tableau simple ; (2) les tableaux complexes (voir Figure 2-B), sont constitués d'en-têtes composées d'une ou plusieurs lignes ou de plusieurs cellules. Plusieurs lignes peuvent être associées à une même cellule d'en-tête. Elles peuvent également

contenir des cellules vides et des cellules fusionnées.

Prenons comme exemple, le tableau (B) de la Figure 2, l'analyse visuelle de ce tableau permet d'identifier les difficultés suivantes :

— les en-têtes des colonnes correspondent généralement aux variables utilisées dans l'étude (par exemple, le nombre de cas, la probabilité, etc) mais elles peuvent figurer sur plusieurs lignes et comportent parfois des cellules vides ;
— les lignes sont composées de cellules qui peuvent parfois être vides. Généralement, le contenu de la première cellule correspond à une variable de l'étude statistique et le reste à la valeur de l'association entre l'en-tête des colonnes et de la ligne ;
— les contenus des cellules peuvent être différents (texte, chiffres, caractères spéciaux, etc.).

L'approche proposée dans cet article prend en compte à la fois la démarche d'analyse d'un tableau statistique en tenant compte des difficultés précédemment identifiées.

## 3 Approche proposée

Nous avons conçu une application qui repose sur l'approche décrite en Figure 3. Le traitement prend en entrée un modèle statistique au format PDF, et se décompose en 3 étapes :

**E1** : extraction de connaissances sous forme de triplets RDF.

**E2** : construction d'un modèle conceptuel.

**E3** : visualisation dynamique du modèle conceptuel.

Dans cet article, nous présentons l'approche permettant la traduction du modèle statistique vers le modèle conceptuel (E1 et E2).

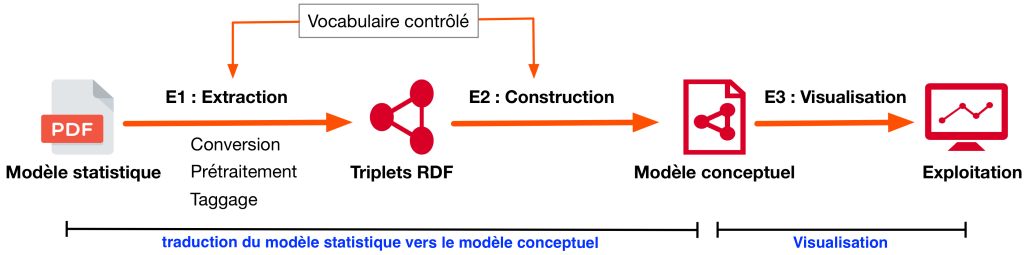



FIGURE *3 – Etapes de la méthodologie générale*

### 3.1 Extraction des connaissances (E1)

Cette section présente un état de l'art relatif à l'extraction de connaissances (3.1.1) ainsi que l'approche d'extraction de connaissances sous forme de triplets RDF à partir du document PDF (3.1.2 et 3.1.3).

#### 3.1.1 Etat de l'art

Ces dernières années, le web est devenu un référentiel universel de données. L'accessibilité de ces données a permis l'émergence de nombreuses approches d'extraction de connaissances à partir de données structurées et non structurées (Unbehauen *et al.*, 2012). Souvent, le flux de données traitées par ces approches d'extraction est issu de tables au sens HTML. L'extraction de ces connaissances permet d'alimenter un grand nombre d'applications (Crestan & Pantel, 2010). La richesse et l'utilité des tables relationnelles sur le web ont permis l'émergence de plusieurs approches d'extraction de connaissances sous forme de triplets RDF (Muñoz *et al.*, 2014), (Lu *et al.*, 2015) ou sous forme d'une description logique (Pivk *et al.*, 2007).

D'autres format sont également utilisés pour coder des tables, le format PDF est le plus répandu. Pour extraire et exploiter ces contenus, il est nécessaire de mettre en place des approches appropriées (Ronzano & Saggion, 2016). Les approches mises en place sont personnalisées et sont fondées sur des éléments structurels tels que le titre, les sections, les figures, les tableaux, etc. (Riaz *et al.*, 2016), (Wu *et al.*, 2015). Cependant, pour exploiter les documents au format PDF, il est nécessaire de les convertir vers un format exploitable (XML, Textuel, HTML, etc.).

Dans cet article, nous nous concentrons sur l'extraction de triplets RDF à partir d'études statistiques publiées au format PDF. Le document PDF est converti dans un format exploitable. Puis, à l'aide d'un vocabulaire contrôlé, les triplets RDF sont extraits à partir des tables HTML.

### 3.1.2 Conversion du fichier PDF

Le format PDF est devenu une norme du support de lecture numérique (ordinateurs, liseuses, tablettes, smartphones, PDA, etc.). L'objectif initial du PDF était de préserver et protéger le contenu et la mise en page d'un document, quels que soient la plate forme ou le programme informatique dans lequel il est visualisé. C'est pourquoi, les fichiers PDF sont difficiles à modifier et parfois même, l'extraction d'information à partir de ces fichiers constitue un véritable défi.

| | **Plateforme** | **Licence** | **Langage de programmation** | **Mise en page conservée** | **Format de sortie** | **Dernière mise à jour** | **Extraction de table** | **Traitement** |
|---|---|---|---|---|---|---|---|---|
| **Tabula** | Application et service web | MIT | Java | Sélection manuelle | JSON CSV | 27/02/2018 | Oui *via* l'intervention de l'utilisateur | Semi-automatique |
| **PDFMiner** | Ligne de commande | MIT | Python | Non | Text, HTML | 24/03/2014 | Non | automatique |
| **Pdftohtml** | Ligne de commande | Copyright | C | Non | HTML, XML | 03/08/2006 | Non | automatique |
| **Pdf2htmlEX** | Ligne de commande | GPLv3 | Python | Oui | HTML | 11/12/2016 | Non | automatique |
| **PDFX** | Service web et ligne de commande | - | JavaScript | Non | HTML,XML | - | Non | automatique |
| **PDF Online** | Service web et Api | Copyright | - | Oui | HTML | - | Non | automatique |

FIGURE *4 – Tableau comparatif des outils de conversion de documents PDF*

En outre, la forme des fichiers PDF varie, ce qui conduit à la mise en place de méthodes de traitements adaptées à chacune d'entre elles. Dans tous les cas, pour automatiser l'extraction d'information, il convient de convertir ces fichiers dans un format exploitable par la machine. Plusieurs outils ont été développés pour aider ce processus de conversion. Pour justifier le choix de l'outil que nous avons utilisé, nous proposons une analyse de ceux souvent cités comme référence :

— pdftohtmlEX [3] convertit les fichiers PDF au format HTML en conservant le texte et la mise en forme des tableaux ;
— pdftohtml [4] convertit les fichiers PDF au format HTML et XML ;
— PDFX [5] utilise des règles pour reconstruire la structure logique d'articles scientifiques au format PDF, quels que soient leur style de formatage ;
— Tabula [6] extrait semi-automatiquement des tableaux de données à partir de fichiers PDF ;
— PDFMiner [7] extrait des informations à partir de documents PDF. Contrairement à d'autres

3. http ://coolwanglu.github.io/pdf2htmlEX/
4. http ://pdftohtml.sourceforge.net/
5. http ://pdfx.cs.man.ac.uk/
6. http ://tabula.technology/
7. http ://www.unixuser.org/ euske/python/pdfminer/

outils liés au PDF, il est entièrement consacré à l'obtention et l'analyse de données de texte ;
— PDF Online [8] convertit les fichiers PDF au format HTML en conservant le texte et la mise en forme des tableaux.

Une synthèse des caractéristiques de ces outils est présentée dans le tableau de la Figure 4. Nous avons sélectionné l'outil pdftohtmlEX (voir Figure 5), pour les raisons suivantes : (1) le format de sortie HTML préserve la structure tabulaire lors du processus de conversion ; (2) l'information concernant la position du tableau dans le document est présente. Ainsi, il est possible de refaire l'action inverse, par exemple en partant des éléments extraits il est possible d'identifier l'endroit du tableau contenant cette information ; (3) la conversion est complètement automatique contrairement à l'outil Tablula. En revanche, la sortie HTML obtenue présente l'inconvénient d'être linéaire et nécessite des traitements supplémentaires.

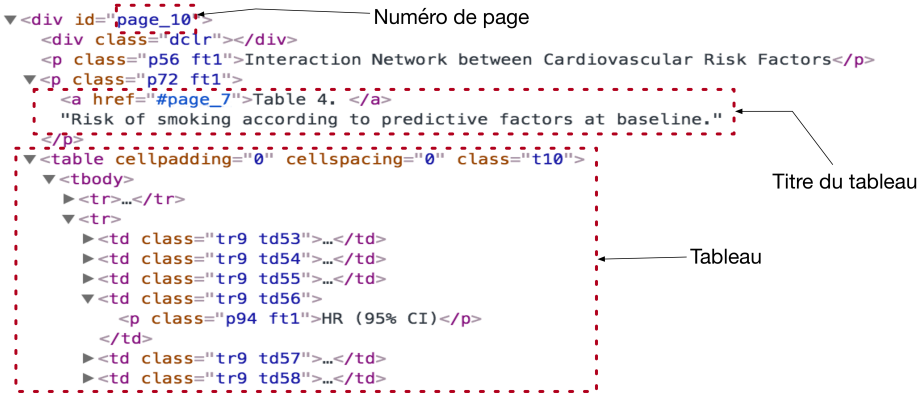



FIGURE 5 – *Sortie HTML obtenue avec pdftohtmlEX*

### 3.1.3 Localisation et extraction de l'information

#### *3.1.3.1 Localisation de l'information*

Une des difficultés de ce travail concerne la reconnaissance de tableaux contenant des informations relatives à un contexte d'étude. Il est facile pour un humain d'identifier les tableaux pertinents dans ce contexte. Par exemple pour un humain, l'analyse visuelle d'un tableau peut lui permettre de déduire facilement le sujet étudié. Ce processus de déduction est réalisé en identifiant certains termes dans le titre, légende, etc. L'automatisation de ce processus est souvent non triviale, même lorsque le système peut localiser les tableaux par reconnaissance de balises « TABLE » dans le document HTML. Le problème à résoudre est : comment déterminer automatiquement si les informations décrites dans un tableau concernent le contexte de l'étude ?

Plutôt que de localiser les informations de manière similaire à (Ermilov *et al.*, 2013) ou de trouver des heuristiques comme l'ont suggéré (Shigarov, 2015) et (Clark & Divvala, 2015), nous avons opté pour une approche supervisée exploitant un vocabulaire contrôlé s'il existe ou à partir des données de l'étude. Nous identifions les chaînes de caractères contenant les termes du vocabulaire. Nous pouvons, par exemple, déterminer qu'un tableau est pertinent s'il contient les termes du vocabulaire dans le titre et/ou le corps du tableau. Contrairement aux deux approches qui supposent un travail important d'élaboration d'heuristiques et des règles, notre démarche est générique et ne nécessite qu'un pré-traitement léger.

8. http ://www.pdfonline.com/

#### *3.1.3.2 Extraction de l'information*

Après avoir identifié les tableaux pertinents (contenant des informations sur le contexte de l'étude) dans le document HTML, l'étape suivante consiste à extraire ces informations et les liens qui leur sont associés.

L'extraction des informations à partir d'un tableau nécessite deux étapes : (1) la détermination des colonnes pertinentes qui est obtenue en utilisant un vocabulaire contrôlé ; (2) la construction des paraphrases d'interprétation des données qui est obtenue en prenant en compte le titre de chaque ligne, les en-têtes des colonnes pertinentes et les cellules se trouvant à l'intersection entre la ligne et l'en-tête.

Notre approche d'extraction comprend quatre étapes :

1. **L'extraction des tableaux pertinents** comprend deux phases : (a) la reconnaissance des tableaux ; (b) la vérification de la pertinence du tableau. La reconnaissance des tableaux est utilisée pour identifier et extraire tous les tableaux présents dans le document en utilisant le format des balises HTML. La vérification de la pertinence des tableaux a pour objectif de déterminer si le tableau extrait traite du sujet d'étude. Le principe consiste à taguer chaque terme du vocabulaire reconnu dans le tableau. Ainsi, s'il contient des termes du vocabulaire, il est déclaré pertinent et stocké sous forme d'un tableau associatif ($numero\_de\_page => titre => contenu$). Dans le cas contraire, il est rejeté.
2. **L'extraction des colonnes pertinentes** consiste à extraire les colonnes dont les en-têtes contiennent les termes du vocabulaire. Cependant, les en-têtes des tableaux peuvent apparaître sur plusieurs lignes avec des en-têtes vides ce qui rend la tâche plus difficile. Pour résoudre ce problème, nous avons commencé par identifier les modèles d'en-têtes utilisés dans les tableaux. Puis, nous avons construit des règles d'extraction pour ces modèles. Par exemple, si l'en-tête de la colonne comporte deux lignes alors la règle de traitement consiste à dupliquer le titre de la première ligne dans la seconde ligne. Souvent les tableaux extraits comportent des cellules vides et des lignes non pertinentes. Un traitement de nettoyage et une mise en forme des tableaux doit par conséquent être appliqué.
3. **Le nettoyage et la mise en forme des tableaux** se déroulent en trois étapes :
   — **substitution des cellules vides** : les cellules vides sont dues à la structuration des tableaux sur plusieurs niveaux. Le traitement permet de recopier le contenu de la cellule précédente dans les cellules vides suivantes d'une même colonne. Pour réaliser ce traitement, on doit considérer le type des données des cellules afin de ne pas introduire de biais. Par conséquent, ce traitement concerne uniquement les cellules contenant des chaînes de caractères.
   — **suppression de lignes non pertinentes** : les lignes non pertinentes proviennent de la structuration des tableaux qui fournissent des informations supplémentaires (telles que, l'année de l'étude, la source, etc.). Le traitement permet de vérifier deux paramètres pour chaque ligne du tableau, la présence des termes du vocabulaire et la présence de plusieurs cellules vides. En combinant ces deux paramètres, les lignes non pertinentes sont automatiquement supprimées.
   — **renommage d'en-tête** : certaines cellules d'en-têtes peuvent être vides ce qui constitue un obstacle à la lecture de l'information. Nous avons construit des règles à l'issue de l'analyse de la structure des tableaux. Par exemple, si le titre de la ligne $i$ du tableau $T$ tient sur deux cellules ($C_1$ et $C_2$) alors la valeur des en-têtes pour $C_1$ et $C_2$ sera respectivement « Class » et « Label ». Le traitement appliqué remplace respectivement l'en-tête vide par les valeurs « Class » et « Label ».

   À l'issue de ces trois étapes, on obtient une collection de tableaux nettoyés et exploitables.
4. **La construction de triplets** : résulte de l'extraction des informations réalisée sur la collection des tableaux $T$. Le résultat obtenu est un ensemble de triplets appelé graphe RDF décrivant l'ensemble des tableaux pertinents.

L'approche choisie est décrite dans la Figure 6 (A). L'automatisation du processus de construction des paraphrases est réalisée de la manière suivante :
Un tableau $T$ est décrit par un ensemble de triplets RDF représentant les lignes le constituant. Chaque ligne $i$ du tableau est considérée comme un nœud blanc noté _ :$x_i$. Une ligne $i$ est décrite par un ensemble de triplets $(s_i, p_k, o_{i,k})$, où :
— $s_i$ : correspond au nœud blanc _ :$x_i$ associé a la ligne $i$ ;
— $p_k$ : correspond à un littéral, à partir du contenu de l'en-tête de la colonne $k$ ;
— $o_{i,k}$ : correspond à un littéral typé, à partir du contenu de la cellule à l'intersection entre la ligne $i$ et la colonne $k$.
Chaque ligne $i$ d'un tableau $T$ comporte un titre. Ce titre peut être présent : (1) sur une colonne, dans ce cas, la première colonne sera décrite par un seul triplet ; (2) sur deux colonnes fusionnées, dans ce cas, la première colonne sera décrite par deux triplets. Les en-têtes des colonnes les plus répandus dans les tableaux statistiques sont la probabilité et le nombre d'individus pour chaque variable étudiée. En partant de ce constat, nous avons construit un modèle générique de triplet guidant leur extraction (voir Figure 6 (B)) . Dans ce modèle, chaque ligne du tableau va comporter un « Label », une « Class », un « Nombre de cas » et une « Probabilité ». Les littéraux typés sont dans un premier temps considérés comme des chaînes de caractères mais seront par la suite décomposés en un « réel » et une « unité » afin éviter une perte d'information sur la donnée.
Puis, chaque titre $F$ d'un tableau $T$ est segmenté pour identifier les éléments décrits et le terme désignant la relation permettant de relier les éléments du titre $F$ au tableau $T$.
Ce traitement est réalisé à l'aide de TreeTagger [9]. A partir de l'élément et de la relation identifiées dans le titre $F$, des triplets sont construits selon le modèle $(s, p, o)$ où :
— $s$ correspond a l'élément identifié dans le titre ;
— $p$ correspond au type de relation entre le titre et le tableau ;
— $o$ correspond au nœud blanc « _ :$x_i$ » pour chaque ligne du tableau $T$.

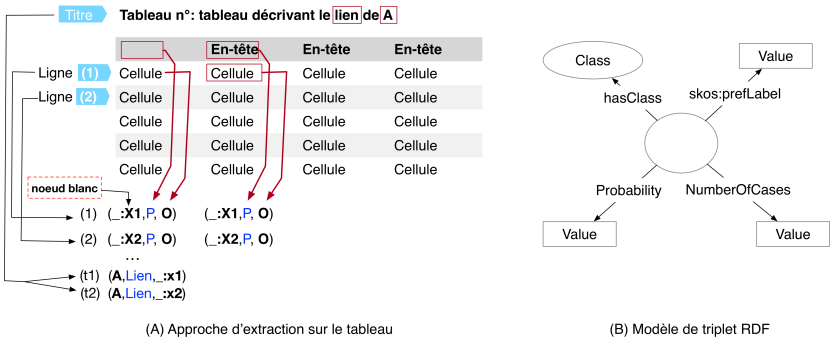



FIGURE 6 – *Stratégie d'extraction des triplets à partir des tableaux*

## 3.2 Construction du modèle conceptuel (E2)

Une des limites des modèles statistiques est qu'ils n'apportent qu'une information brute et faiblement structurée par rapport au contexte d'étude. En effet, la description et l'interprétation d'une relation dans un tableau sont construites sur des observations soumises à la subjectivité de l'observateur. Par exemple, deux personnes ayant des niveaux de connaissances différents sur un sujet n'auront pas la même interprétation d'un tableau.
L'approche proposée pour réduire ce biais consiste à construire un modèle conceptuel du domaine étudié pour guider l'interprétation des connaissances implicites dans les tableaux. L'avantage de ce modèle est : (i) de permettre d'avoir une description structurée ; (ii) d'être souple en permettant d'étendre à d'autres concepts du domaine.

9. http ://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/

## 4 Expérimentation

Ce travail s'inscrit dans le contexte de l'amélioration de la prévention du risque cardiovasculaire qui cause près de 17,7 millions de décès, soit 40% de la mortalité mondiale totale (WHO, 2017). Si les principaux facteurs de risques cardiovasculaires sont aujourd'hui bien connus, leur évaluation à tendance à être réalisée sans considérer les interactions qui les lient (Meneton *et al.*, 2016).

Pour conduire notre expérimentation, nous travaillons à partir d'une étude statistique menée par (Meneton *et al.*, 2016) dans le domaine de l'épidémiologie. Son objectif était de mettre en évidence des interactions entre des facteurs de risques cardiovasculaires. Les résultats obtenus par les auteurs en appliquant des tests de régression sont résumés dans des tableaux, figures et textes.

Le but de notre expérimentation est double : extraire les connaissances contenues dans les tableaux sous forme de triplets RDF et construire le modèle conceptuel à partir de ces triplets.

L'étude de (Meneton *et al.*, 2016) est publiée sous forme d'un document PDF dans lequel 13 tableaux décrivent les associations de 13 facteurs de risque cardiovasculaires. La première étape était de convertir le fichier PDF vers un format exploitable à l'aide de pdftohtmlEX. Le résultat obtenu est un document HTML contenant « 9798 lignes », « 55702 mots », « 1254107 caractères » et « 32 tableaux ».

Pour illustrer notre démarche, nous avons choisi un tableau (voir Figure 7) du document PDF. L'ensemble des tableaux décrivant les associations entre les facteurs de risque cardiovasculaires ont tous la même structure.

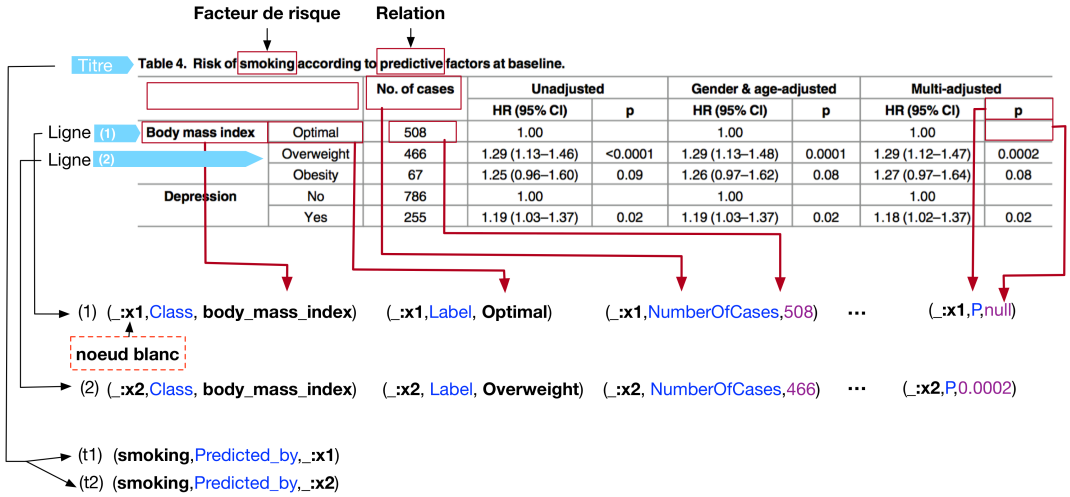
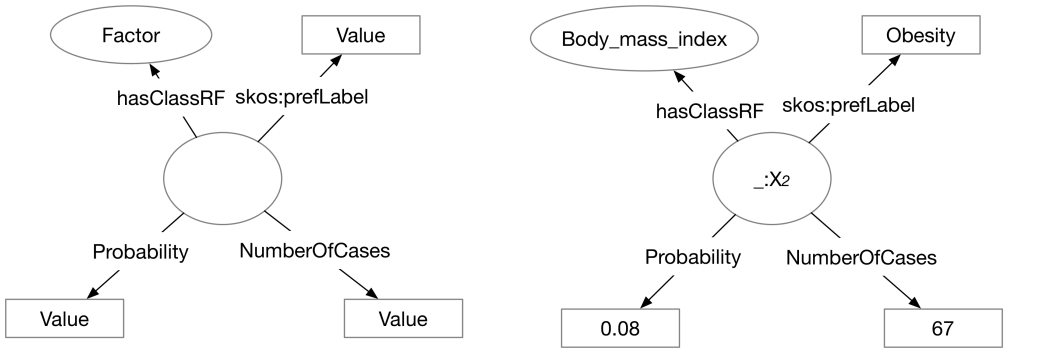



FIGURE 7 – *Stratégie d'extraction de l'information appliquées sur le tableau*

Chaque ligne $i$ du tableau (Figure 7) comporte un titre. Ce titre est écrit sur deux colonnes fusionnées. Le titre de la première ligne du tableau est composé de « Body mass index » et « Optimal ». Dans ce cas, la première colonne sera décrite par deux triplets de la forme $(\_ :\mathrm{x_i}, p, o)$, où :

— « $\_ :\mathrm{x_i}$ » correspond à la valeur du nœud blanc ;
— « $p$ » correspond respectivement à « Class » et « Label » pour la première colonne ;
— « $o$ » correspond à la valeur de l'intersection entre la ligne $i$ et la colonne $k$ du tableau $T$. Les contenus des colonnes « P » (décrivant la probabilité) et « No. of cases » (décrivant le nombre d'individus pour chaque Class) sont traités comme des chaînes de caractères.

Chaque ligne du tableau est décrite par un triplet suivant le modèle décrit dans la Figure 8-A . Par exemple, le résultat obtenu pour la troisième ligne du tableau (A) est présenté Figure 8-B.



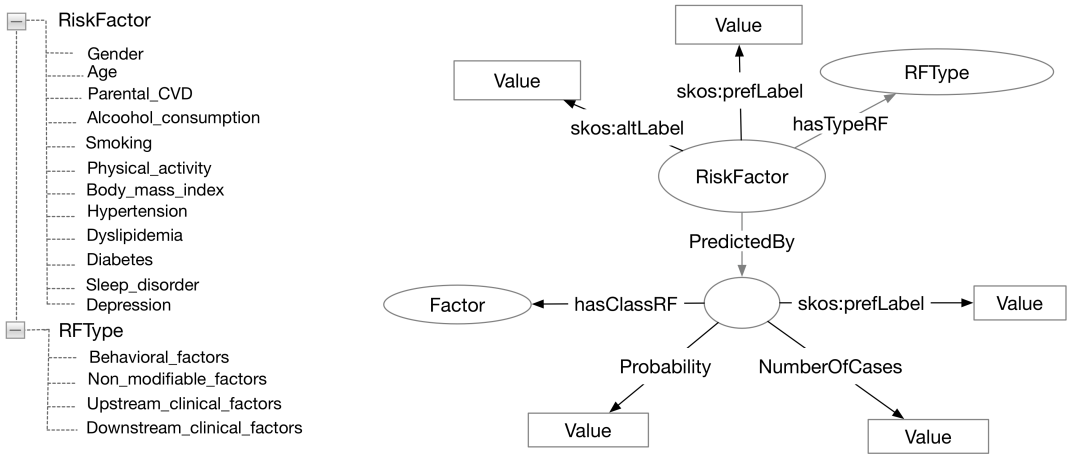

(A) Modèle de triplet

(B) Triplet obtenu pour la 3ème (_:X2) ligne du tableau (A)

FIGURE *8 – Modèle de triplet pour l'extraction d'information*

— **Construction du modèle conceptuel (E2)** :
«CARRE [10] » est à notre connaissance le seul vocabulaire actuellement disponible. Il décrit les facteurs de risque cliniques cependant, il contient peu de termes relatifs aux facteurs de risque cardiovasculaires et est difficilement réutilisable. Pour pallier cette incomplétude, nous avons élaboré un vocabulaire contrôlé en utilisant les termes du domaine des maladies cardio-vasculaires utilisés par les experts du domaine. Le modèle conceptuel obtenu (voir Figure 9-(2)) a été élaboré en utilisant ce vocabulaire contrôlé. Dans ce modèle, chaque facteur de risque est décrit par sa catégorie, les facteurs qui le prédisent, le label préféré et le label alternatif. Par exemple, le facteur de risque « Smoking » est décrit par :

1. catégorie des « Behavioral_factors » ;
2. prédit par les facteurs associés aux nœuds blancs ($\_ :x_1, \_ :x_2, \_ :x_3$) ;
3. label préféré « Smoking » et label alternatif « Fumeur ».




(1) Vocabulaire contrôlé

(2) Modèle conceptuel

FIGURE *9 – Structure du vocabulaire contrôlé*

10. http ://bioportal.bioontology.org/ontologies/CARRE

— **Localisation de l'information** : le résultat obtenu à partir du tableau (A) est présenté Figure 10. Nous avons identifié à l'aide du vocabulaire contrôlé les termes «Smoking, predictive » dans le titre et « Obesity, Depression » dans le corps du tableau. Ainsi, le tableau (A) est annoté comme pertinent.

```
<P class="p72 ft1"><A href="#page_7">Table 4. </A>Risk of smoking according to predictive factors at baseline.</P>
<TABLE cellpadding=0 cellspacing=0 class="t10">
...
<TR>
      <TD class="tr0 td61"><P class="p0 ft0"> </P></TD>
      <TD class="tr0 td54"><P class="p62 ft30">Obesity</P></TD>
      <TD class="tr0 td55"><P class="p62 ft31">67</P></TD>
      <TD class="tr0 td56"><P class="p81 ft31">1.25 <NOBR>(0.96–1.60)</NOBR></P></TD>
      <TD class="tr0 td57"><P class="p0 ft0"> </P></TD>
      <TD class="tr0 td58"><P class="p88 ft1">0.09</P></TD>
      <TD class="tr0 td56"><P class="p60 ft31">1.26 <NOBR>(0.97–1.62)</NOBR></P></TD>
      <TD class="tr0 td59"><P class="p62 ft31">0.08</P></TD>
      <TD class="tr0 td56"><P class="p81 ft31">1.27 <NOBR>(0.97–1.64)</NOBR></P></TD>
      <TD class="tr0 td25"><P class="p81 ft1">0.08</P></TD>
</TR>
...
</TABLE>
```

FIGURE *10 – Exemple de tag réalisé sur un tableau à l'aide du vocabulaire contrôlé*

— **Extraction des tableaux pertinents** : une fois le tableau (A) déclaré pertinent, il est extrait sous forme d'un tableau associatif (voir Figure 11) « $numero\_de\_page \Rightarrow titre \Rightarrow contenu$ ».

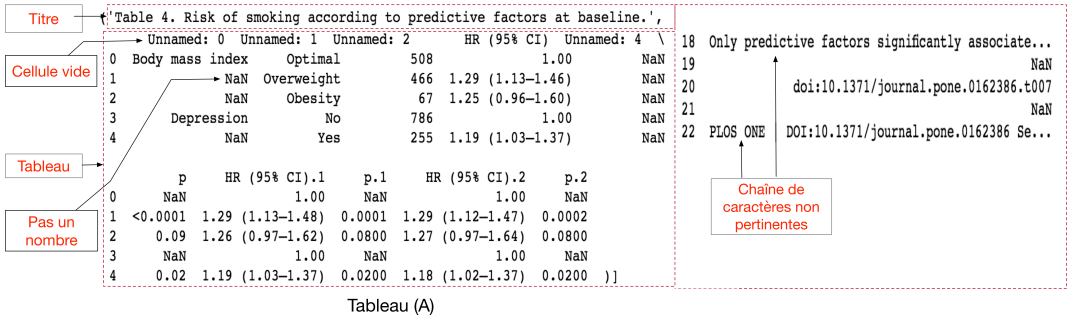



FIGURE *11 – Exemple de tableau extrait à partir du document HTML*

— **Extraction des colonnes pertinentes** : l'extraction des colonnes a été réalisée à l'aide du vocabulaire contrôlé et du modèle de triplet. Les colonnes extraites à partir du tableau (A) sont : (1) la probabilité décrite par l'en-tête portant l'étiquette « P » ; (2) le nombre de cas décrit par l'en-tête portant l'étiquette « No.of cases ».

('Table 4. Risk of smoking according to predictive factors at baseline.',

| | Facteur | Label | NumberOfCases | Probability |
|---|---|---|---|---|
| 1 | Body_mass_index | Overweight | 466 | 0.0002 |
| 2 | Body_mass_index | Obesity | 67 | 0.0800 |
| 3 | Depression | Yes | 255 | 0.0200 |

FIGURE *12 – Tableau (A) de la Figure 11 après nettoyage*

— **Nettoyage et mise en forme du tableau** : les résultats obtenus pour le tableau (A) sont présentés dans la Figure 12. Nous avons constaté que certains champs de la colonne «

probability » comporte des valeurs « NaN ». Cette valeur traduit la non-pertinence de la ligne. Nous avons appliqué un traitement de suppression de toutes les lignes contenant la valeur « NaN » dans la colonne « probability » pour l'ensemble des tableaux.

— **Construction de triplets** : le résultat obtenu pour le tableau (A) est présenté dans la Figure 13. Nous avons obtenu cinq nœuds blancs notés (_ :$x_1$, _ :$x_2$, _ :$x_3$). En prenant l'exemple du nœud blanc _ :$x_1$ associé à la première ligne du tableau, cinq triplets sont produits. Quatre triplets construits à partir de la ligne « 1 » et des colonnes « $Probability$ » et « $NumberOfCases$ » qui sont :
(_ :$x_1$,'$Class$','$Body\_mass\_index$'), (_ :$x_1$,'$Label$','$Overweight$'),
(_ :$x_1$,'$NumberOfCases$','466'), (_ :$x_1$,'$Probability$','0.0002'). Le cinquième triplet construit à partir du titre $T$ et de la ligne « 1 » est ('$Smoking$','$PredictedBy$', _ :$x_1$).

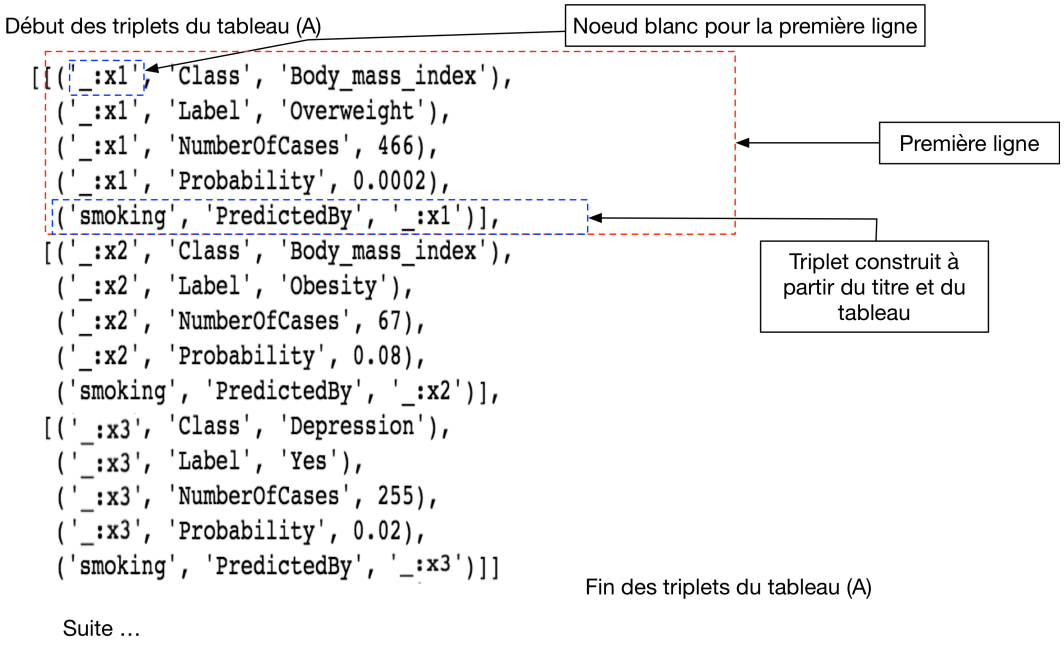



FIGURE *13 – Triplets obtenus à partir du tableau (A) de la Figure 11*

A l'issue de l'extraction des triplets, chaque facteur de risque est décrit selon le modèle conceptuel présenté dans la Figure 9-(2). Ce processus est réalisé en deux étapes : (1) analyser les triplets extraits ; (2) associer à chaque facteur de risque un ensemble de triplets.

Le résultat est stocké dans un Triple Store. Ainsi, le langage de requête SPARQL peut être utilisé directement pour interroger ce graphe. Outre son interrogation, le modèle obtenu est destiné à être utilisé dans une nouvelle approche d'évaluation du risque cardiovasculaire fondée sur la visualisation dynamique des interactions entre les facteurs de risque.

## 5 Evaluation

Nous avons montré dans la section précédente comment extraire et transformer des connaissances à partir des tableaux statistiques au format PDF. L'approche développée est adaptable à d'autres formats et à usage général. Elle est réalisée en deux étapes : (1) localisation et extraction de l'information dans les tables ; (2) élaboration du modèle conceptuel. Ces deux étapes sont fondées sur l'utilisation d'un vocabulaire contrôlé.

Afin de valider notre approche, nous avons conduit une expérimentation fondée sur l'interprétation par un expert de la sélection de tableaux statistiques correspondant à un sujet d'étude dans un document PDF. Pour évaluer l'approche d'extraction des tableaux, nous avons travaillé sur deux jeux de données disponibles au format PDF : le premier (D1) concerne le modèle statistique à l'origine de ce travail dans le domaine des maladies cardiovasculaires, la

démarche et les résultats obtenus sont présentés dans la section 4 ; le second (D2) concerne l'enquête internationale sur les transactions de change et de produits dérivés [11] dans le domaine financier. La Figure 14 présente l'exemple d'un tableau extrait du document et le résultat obtenu après extraction des triplets est présenté dans la Figure 15.

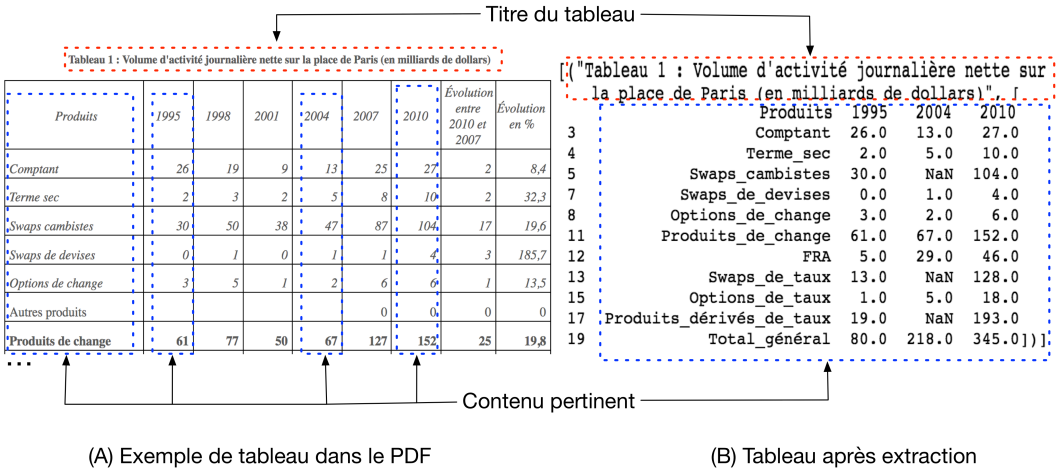



FIGURE *14 – Extraction d'un tableau figurant dans (D2)*

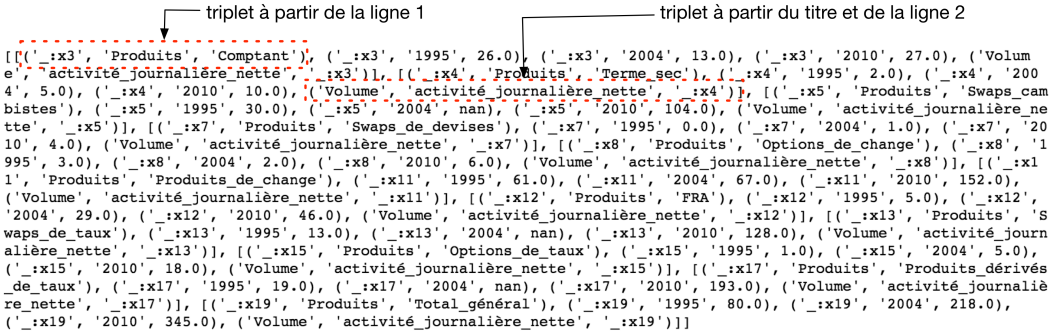



FIGURE *15 – Liste des triplets obtenus à partir du tableau de la Figure 14(B)*

Nous avons utilisé trois métriques pour évaluer les résultats de l'extraction sur (D1) et (D2) : la *précision*, le *rappel* et la *F-mesure*. Soit $T_{fac}$ le nombre de tableaux identifié comme traitant des interactions entre les facteurs de risque cardiovasculaire (D1) et au volume d'activité de transactions (D2) ; la précision est le rapport entre $T_{fac}$ et le nombre total de tableaux apparaissant dans chaque jeu de données ; le rappel est le rapport entre $T_{fac}$ et les tableaux décrivant réellement les interactions entre les facteurs de risque cardiovasculaire (D1) et le volume d'activité de transactions (D2) ; la F-mesure est la moyenne quadratique combinant la précision et le rappel. Pour chaque jeu de données, nous avons exécuté une extraction avec et sans vocabulaire contrôlé. Une fois l'extraction réalisée, nous avons calculé la *précision*, le *rappel* et la *F-mesure*.

La Table 1 présente les résultats de l'extraction des tableaux avec les mesures de *précision*, *rappel* et *F-mesure* pour les jeux de données (D1) et (D2). Nous constatons que la précision

11. https ://www.banque-france.fr/sites/default/files/media/2016/11/24/enquete-triennale-principaux-resultats.pdf

| Jeux de données | Vocabulaire | Précision | Rappel | F-mesure |
|---|---|---|---|---|
| D1 | sans | 0.40 | 0.90 | 0.55 |
| | avec | **1.0** | **0.98** | **0.99** |
| D2 | sans | 0.50 | 0.45 | 0.47 |
| | avec (1 terme) | **0.76** | **0.69** | **0.72** |
| | avec (2 termes ) | **0.89** | **0.80** | **0.84** |

TABLE *1 – Résultats expérimentaux*

de l'extraction donne de faibles résultats sur les deux jeux de données lorsqu'elle est réalisée sans le recours à un vocabulaire. Cette faible valeur de la précision est due au nombre élevé de tableaux extraits. La *précision* et la *F-mesure* sur (D1) augmentent lorsqu'un vocabulaire contrôlé est utilisé. Pour le jeu de données (D2), la *précision* et le *rappel* augmentent au fur et à mesure où le nombre de termes augmente dans le vocabulaire. Ces résultats indiquent que l'utilisation d'un vocabulaire adapté est déterminant pour optimiser l'extraction.

L'évaluation de cette approche sur le jeux de données (D1) fournit une performance presque parfaite pour le jeu de données (D1). Ce résultat s'explique principalement par : (1) l'approche développée sur le jeu de données (D1) ; (2) les tableaux décrivant les interactions entre les facteurs de risque cardiovasculaires dans le document HTML ont tous la même structure ; (3) le vocabulaire utilisé pour l'extraction est adapté au domaine. Sur le jeu de données (D2), la performance reste très satisfaisante. Nous prévoyons de tester l'approche d'extraction sur un volume plus important de jeux de données pour mieux évaluer l'approche et identifier des pistes d'amélioration.

Le cadre proposé dans cet article n'est pas limité à l'extraction de connaissances à partir d'études statistiques au format PDF, mais peut être appliqué à toutes ressources structurées sous forme de tableaux. L'originalité de cette approche est d'associer un modèle conceptuel aux tableaux figurant dans un document PDF.

Une autre expérimentation, en cours avec les chercheurs en statistiques, montre la difficulté de l'interprétation des connaissances représentées dans le modèle statistique. Les premiers résultats sont encourageants, ils démontrent outre un gain de temps, l'apport du langage SPARQL qui facilite l'accès aux connaissances (par exemple, filtrage sur la probabilité, le nom de facteur de risque, etc.).

## 6 Conclusion et perspectives

Dans cet article, nous avons décrit une méthode permettant la traduction d'un modèle statistique présenté sous forme de tableau et publié au format PDF, vers un modèle conceptuel représenté sous la forme d'un graphe. Nous avons apporté des solutions à deux problèmes dans le domaine de l'extraction de connaissances : (i) comment déterminer la pertinence des informations contenues dans des tableaux et sous quel format les extraire ; (ii) comment passer d'un format PDF non structuré à un format exploitable pour le traitement sémantique de l'information. Une réponse au second problème est constituée de la conversion d'un document PDF vers un format HTML respectant la structure des tableaux, puis l'extraction des informations pertinentes sous forme de triplets RDF. L'intérêt de cette approche est qu'elle permet d'extraire des connaissances implicites représentées dans des tableaux statistiques dans différents domaines.

Les résultats de nos premières expérimentations sur des ensembles de données de nature différentes sont encourageants, même s'ils doivent encore être améliorés. Plusieurs perspectives émergent comme l'ajout de l'exploitation du contenu complet du document (texte, figure, etc.). Le résultat obtenu est déjà intégré dans un système de visualisation [12] dynamique de connaissances appliqué aux interactions entre les facteurs de risque des maladies cardiovasculaires. En outre, l'approche est actuellement utilisée sur des études statistiques dans

12. http ://www-limics.smbh.univ-paris13.fr/MCVGraphViz/

le domaine des maladies cardiovasculaires et conduit à des modèles conceptuels différents. L'idée est de fusionner ces modèles en exploitant les connaissances expertes du domaine afin d'élaborer un modèle générique des interactions entre les facteurs de risque des maladies cardiovasculaires.

## Références


CLARK C. A. & DIVVALA S. K. (2015). Looking beyond text : Extracting figures, tables and captions from computer science papers. In *Scholarly Big Data : AI Perspectives, Challenges, and Ideas, Papers from the 2015 AAAI Workshop, Austin, Texas, USA, January, 2015.*

CRESTAN E. & PANTEL P. (2010). Web-scale knowledge extraction from semi-structured tables. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, p. 1081–1082, New York, NY, USA : ACM.

ERMILOV I., AUER S. & STADLER C. (2013). User-driven semantic mapping of tabular data. In *Proceedings of the 9th International Conference on Semantic Systems*, I-SEMANTICS '13, p. 105–112, New York, NY, USA : ACM.

JUNYONG I. & SANGSEOK L. (2017). Statistical data presentation. *Korean Journal of Anesthesiology*, **70**, 267.

KLAMPFL S. & KERN R. (2015). Machine learning techniques for automatically extracting contextual information from scientific publications. In *Semantic Web Evaluation Challenges*, p. 105–116. Springer International Publishing.

LU W., ZHANG Z., LOU R., DAI H., YANG S. & WEI B. (2015). Mining rdf from tables in chinese encyclopedias. In *Proceedings of the 4th CCF Conference on Natural Language Processing and Chinese Computing - Volume 9362*, NLPCC 2015, p. 285–298, Berlin, Heidelberg : Springer-Verlag.

MENETON P., LEMOGNE C., HERQUELOT E., BONENFANT S., LARSON M.-G., VASAN R.-S., MÉNARD J., GOLDBERG M. & ZINS M. (2016). A global view of the relationships between the main behavioural and clinical cardiovascular risk factors in the gazel prospective cohort. *PLOS ONE*, **11**(9), 1–20.

MUÑOZ E., HOGAN A. & MILEO A. (2014). Using linked data to mine rdf from wikipedia's tables. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, p. 533–542, New York, NY, USA : ACM.

PIVK A., CIMIANO P., SURE Y., GAMS M., RAJKOVIČ V. & STUDER R. (2007). Transforming arbitrary tables into logical form with tartar. *Data Knowl. Eng.*, **60**(3), 567–595.

POWELL J. (2015). A librarian's guide to graphs, data and the semantic web. Chandos Information Professional Series, p. 268. Elsevier Science.

RIAZ A., TANVIRAND A. M. & MUHAMMAD A. Q. (2016). Information extraction from PDF sources based on rule-based system using integrated formats. In *Semantic Web Challenges*, p. 293–308. Springer International Publishing.

RONZANO F. & SAGGION H. (2016). Knowledge extraction and modeling from scientific publications. In A. GONZÁLEZ-BELTRÁN, F. OSBORNE & S. PERONI, Eds., *Semantics, Analytics, Visualization. Enhancing Scholarly Data*, p. 11–25, Cham : Springer International Publishing.

SAPORTA G. (2011). Probabilités, analyse des données et statistique. p. 622. 3ème édition révisée.

SHIGAROV A. O. (2015). Table understanding using a rule engine. *Expert Systems with Applications*, **42**(2), 929–937.

UNBEHAUEN J., HELLMANN S., AUER S. & STADLER C. (2012). *Knowledge Extraction from Structured Sources*, In S. CERI & M. BRAMBILLA, Eds., *Search Computing : Broadening Web Search*, p. 34–52. Springer Berlin Heidelberg : Berlin, Heidelberg.

WHO (2017). *World Health Statistics 2017 :Monitoring Health for the SDGs Sustainable Development Goals*. World Health Statistics Annual. World Health Organization.

WU J., KILLIAN J., YANG H., WILLIAMS K., CHOUDHURY S. R., TUAROB S., CARAGEA C. & GILES C. L. (2015). Pdfmef : A multi-entity knowledge extraction framework for scholarly documents and semantic search. In *Proceedings of the 8th International Conference on Knowledge Capture*, K-CAP 2015, p. 13 :1–13 :8, New York, NY, USA : ACM.

YEON-SEOK K. & KYONG-HO L. (2008). Extracting logical structures from HTML tables. *Computer Standards & Interfaces*, **30**(5), 296–308.

# De l'intérêt des ontologies modulaires. Application à la modélisation de la prise en charge de la SLA

Sonia Cardoso[1,3], Xavier Aimé[2,3], Vincent Meininger[4], David Grabli[5], Kevin Bretonnel Cohen[6,7], Jean Charlet[3,8]

[1] IHU-A-ICM, Hôpital Pitié Salpêtrière
sonia.cardoso@icm-institute.org

[2] Cogsonomy

[3] Sorbonne Université, INSERM, Université Paris 13, Sorbonne Paris Cité, UMR_S 1142, LIMICS, Paris

[4] Ramsay General de Santé, Hôpital Peupliers, Paris, France

[5] AP-HP Pitié Salpêtrière, Département des maladies du Système Nerveux, Paris SU, France

[6] Computational Bioscience Program, U. Colorado School of Medicine, USA

[7] Université Paris-Saclay, LIMSI-CNRS, France

[8] Assistance Publique-Hôpitaux de Paris, DRCI, Paris, France

**Résumé** : Pour tenter de comprendre les causes de ruptures de parcours de soins dans le cadre de la prise en charge des patients atteints de Sclérose Latérale Amyotrophique (SLA), dans un réseau de coordination de parcours de soins, nous avons crée une ontologie modulaire. L'objectif de notre travail est d'expliciter les apports et limites d'une ontologie modulaire, dans ce cas d'usage, la méthodologie de construction utilisée et de la comparer d'un point de vue quantitatif à des ontologies monolithiques de domaines connexes.

**Mots-clés** : Ontologie, ontologie modulaire, parcours de soin, Sclérose Latérale Amyotrophique.

## 1 Introduction

La sclérose latérale amyotrophique (SLA) est une maladie neurodégénérative qui entraîne une détérioration rapide et progressive du contrôle volontaire de la musculature squelettique et éventuellement une perte de fonction des muscles de la ventilation, entraînant la mort entre 24 et 36 mois (Corcia *et al.*, 2008). Selon les pays, le suivi et la prise en charge se fait différemment, par exemple, aux États-Unis, la prise en charge de la SLA se fait principalement à domicile (Lavernhe *et al.*, 2017). En France, la prise en charge de cette maladie s'oriente vers le domicile mais elle peut néanmoins provoquer des hospitalisations fréquentes et des retours ultérieurs dans le milieu familial. À Paris, fut créé un réseau de coordination : le réseau de SLA dont l'objectif principal est de coordonner la prise en charge des patients atteints de SLA. La paralysie progressive des muscles va placer les patients face à de nombreuses situations de handicap. Les patients et leur entourage vont nécessiter différents types d'aides, des aides humaines lors de la réalisation d'activités de vie quotidienne – manger, se laver, s'habiller, etc. – des aides techniques (fauteuil roulant électrique ou manuel) pour les déplacements notamment, mais aussi un accompagnement social pour la mise en place de financement des aides (Soriani & Desnuelle, 2017). Des problèmes surgissent lorsque les transitions entre l'hôpital et le lieu de vie entraînent des interruptions dans le parcours de soin. Ces ruptures peuvent avoir un impact sur la santé et la qualité de vie du patient et de sa famille. Cependant, nous ne disposons d'aucune donnée sur l'origine des problèmes entrainant ces ruptures de parcours à domicile. Ce manque de compréhension des causes des ruptures est un problème car, si les causes des interruptions de soins étaient comprises, des mesures pourraient être prises pour les prévenir.

Un des moyens de mieux appréhender ces causes de rupture de soins et de les comprendre, consisterait entre autres à pouvoir les modéliser pour mieux les analyser. Une ressource permettant cette modélisation existe au travers d'un corpus provenant d'un système de messagerie. Ce système est utilisé pour la communication entre les coordonnateurs du réseau SLA et les personnes impliquées dans la prise en charge des patients (professionnels de santé de proximité, membres de la famille, structure sociale et médico-sociale et patients). Les données sont des messages sous forme textuelle exprimés en langage naturel mais non structurés, ce qui les rend intelligibles aux humains mais non analysables par les techniques classiques d'exploration de données. De plus, la quantité des messages est très importante, ce qui rend inenvisageable de les analyser manuellement – un cas d'utilisation classique pour le traitement automatique du langage naturel (Lin, 2008).

Afin de faciliter l'exploration de texte et le travail d'exploration de données subséquent, nous envisageons d'utiliser une ontologie. Cependant, aucune ontologie incluant les dimensions médicale, de coordination et socio-environnementale du système social et médico-social français n'est disponible ou n'a été créée à ce jour. Nous avons donc été obligé de construire une telle ontologie, ONTOPARON, dans le but de mieux comprendre les causes de rupture de soins dans le cadre de la SLA. C'est une tâche spécifique car elle nécessite une modélisation de la communication (entre les professionnels et les membres de la famille des patients), alors que les ontologies biomédicales sont généralement orientées vers la modélisation de la biologie et de la pathologie. Pour cela, nous avons adopté une forme modulaire de l'ontologie.

Nous axons cet article sur le caractère modulaire de l'ontologie et les difficultés de sa construction. Nous exposerons les motivations pour la conception modulaire et le travail de modularisation lui-même. Nous aborderons les difficultés que nous avons rencontrées et nous discutons les points qui nous paraissent saillants dans ce travail. Nous ne nous préoccuperons pas des problèmes formels et de langages de représentation associés aux ontologies et aux ontologies modulaires comme ils sont décrits dans (Bao & Honavar, 2006) et nous considérerons, ce qui est notre cas, que le langage de représentation OWL utilisé est suffisamment expressif pour représenter les connaissances nécessaires.

## 2 Définition et motivation de la modularité d'une ontologie

Pathak *et al.* (2009) définit une ontologie modulaire comme un ensemble de modules qui sont des « composants réutilisables d'une ontologie plus grande ou plus complexe, qui est autonome mais qui présente une association définie avec d'autres modules d'ontologie, y compris l'ontologie originale ». En ce sens, les ontologies modulaires contrastent avec les ontologies monolithiques.

Selon Bao et Honavar (2006), « comme les ontologies sont conçues pour des domaines spécifiques, des applications, ou des utilisateurs, elles nécessitent souvent des adaptations importantes avant de pouvoir être déployées avec succès dans des environnements connexes et proches. Il y a donc un besoin urgent de favoriser une réutilisation sélective et indépendante de modules au sein de ces ontologies. » D'une manière générale, « les ontologies modulaires (1) facilitent la réutilisation de connaissances sur de multiples applications, (2) sont faciles à construire, maintenir et modifier, (3) permettent une ingénierie distribuée des modules sur différents champs d'expertise, et (4) permettent une gestion et une navigation efficace dans les modules » (Grau *et al.*, 2006).

Mais par rapport aux ontologies monolithiques, les ontologies modulaires posent des défis supplémentaires. Par exemple, elles imposent une charge supplémentaire de spécification d'un niveau supplémentaire dans la hiérarchie d'héritage pour chaque concept.

Abbès *et al.* (2012) ont étudié les caractéristiques de la modularité et ont proposé de classer la modularisation en quatre types de patrons : 1) 1 module important n modules, 2) n modules important 1 module, 3) n modules important n-1 modules et, enfin, 4) un mixte de tous ces patrons. La figure 1 précise les deux premiers patrons pour les importations. Ils sont les plus importants pour la construction modulaire puisqu'ils représentent, respectivement, la modularité de l'ontologie via l'agrégation de plusieurs modules et l'héritage d'un module correspondant la plupart du temps à une ontologie noyau.

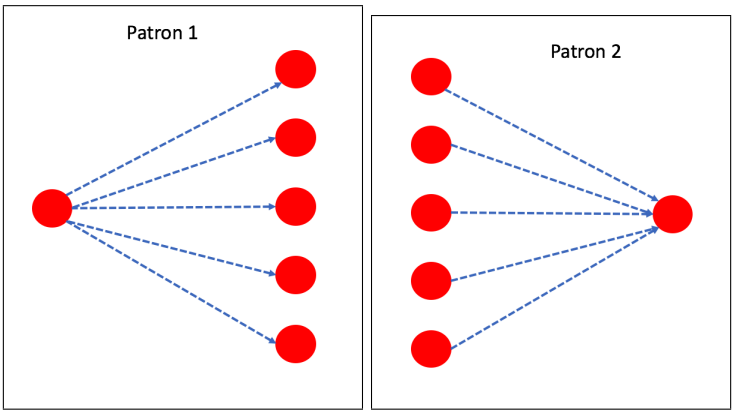



FIGURE *1 – Premier et second patrons de type d'importation de modules. Le premier patron correspond à une action d'agrégation où l'on construirait une ontologie des animaux en agrégeant des ontologies des mammifères, des oiseaux, etc. Le second schéma correspond à une action d'héritage où n modules se partagent les concepts d'un module plus général. C'est typique des actions de partage d'ontologie noyau. Schémas tirés de Abbès* et al. *(2012).*

## 3 Construire une ontologie modulaire

### 3.1 Éléments de modularisation

En pratique, quand on veut construire une ontologie modulaire, on se trouve devant deux cas :

1. on part de zéro (pas d'ontologie monolithique ou modulaire déjà construite) et on décide de construire une ontologie modulaire ;
2. on est déjà dans un contexte d'ontologie modulaire. En conséquence, le module à construire doit s'inscrire parfaitement dans ce contexte. Et l'ontologue se doit de maîtriser ce contexte ; à savoir $a$) connaître les autres modules, *i.e.* les autres ontologies, $b$) y compris les ontologies importées par les autres modules.

Les réflexions que nous retranscrivons ici correspondent au 2$^{e}$ cas et c'est cela que nous avons expérimenté dans ONTOPARON. Dans notre cas, comme précisé en introduction, nous avons constitué trois modules pour les dimensions médicale, de la coordination et socio-environnementale. Deux autres modules sont imposés par la modélisation : (1) l'ontologie noyau qui correspond aux concepts communs aux trois modules (et qui est importée par chacun d'eux) et (2) le module d'agrégation qui met ensemble les trois modules. Ce dernier module ne modélise aucun concept, il ne sert qu'à agréger. La figure 2 schématise les liens d'importation de nos principaux modules.

Nous choisissons d'importer dans chaque module uniquement des modules sémantiquement plus généraux. Même si, techniquement, un éditeur comme Protégé autorise n'importe quel import – quel que soit le niveau de granularité sémantique. Nous estimons en effet que l'importation d'un module se calque sur la notion d'héritage telle que développée dans la programmation orientée objet par exemple. Ainsi une commande *import* dans une ontologie se comporte comme l'instruction *extends* dans une classe en Java. Le module hérite de l'ensemble des concepts et relations qu'il va étendre/spécifier (et non l'inverse).

### 3.2 Processus de construction

La structure modulaire globale de l'ontologie, ainsi que les concepts spécifiques qu'elle contient, ont été développés et validés par un processus itératif combinant une analyse ascendante des données sur la communication liée à la coordination pour les patients SLA et

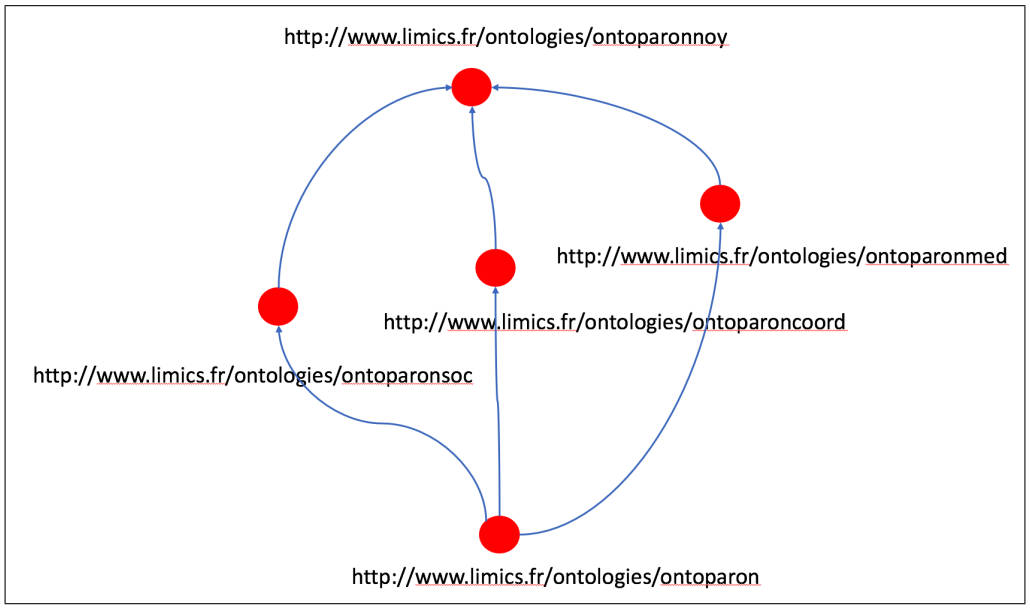



FIGURE 2 – *Vue d'ensemble du schéma d'héritage des concepts d'*ONTOPARON. *Les classes de chacune des ontologie ont une URI spécifique. Les ontologies correspondant aux annotations de l'ontologie et des concepts ne sont pas représentées (*e.g. *SKOS ou* Dublin Core*).*

la modélisation descendante par un expert du domaine. La construction et l'évaluation des ontologies sont des tâches complexes, avec de nombreuses approches valables. Comme suggéré par le travail de Spasic *et al.* sur la mise à jour des réseaux sémantiques par typage sémantique des termes dans un corpus textuel (2005) et par le travail de Friedman *et al.* sur l'évaluation des ontologies via la couverture des informations textuelles (2006), ONTOPARON a été amorcée par l'analyse des messages du corpus. L'outil HeTOP [1] a été utilisé pour aligner les concepts de nos ontologies avec d'autres terminologies en utilisant leurs codes UMLS.

### 3.3 Résultats quantitatifs

Pour faire une analyse quantitative de ONTOPARON, nous la comparons à deux ontologies de domaines apparentées, l'ontologie de coordination des soins infirmiers (NCCO — Popejoy *et al.* (2014)) et l'ontologie de la maladie d'Alzheimer et des maladies connexes (ONTOAD–Dramé *et al.* (2014)). La première est pertinente dans la mesure où son domaine est lié à celui de la coordination des soins, la seconde, dans la mesure où elle modélise une maladie dégénérative neurologique. Puisque ONTOAD a été développée en français et en anglais, comme ONTOPARON, elle permet également de comparer les aspects bilingues de notre ontologie.

Pour faire une analyse quantitative de ONTOPARON, nous la comparons à deux ontologies de domaines apparentées : l'ontologie de coordination des soins infirmiers (NCCO) qui est pertinente dans la mesure où son domaine est lié à celui de la coordination des soins ; l'ontologie de la maladie d'Alzheimer et des maladies connexes (ONTOAD) qui est pertinente dans la mesure où elle modélise une maladie dégénérative neurologique. Puisque ONTOAD a été développée en français et en anglais, comme ONTOPARON, elle permet également de comparer les aspects bilingues de notre ontologie.

Le tableau 3 quantifie quatre aspects d'OntoParon et des deux ontologies connexes : le nombre de classes, le nombre de termes en anglais et en français, et la profondeur médiane des concepts.

1. `https://hetop.eu/hetop/`

Nombre de classes : la quantité de classes est une estimation de la couverture du domaine. Un grand nombre de classes permet une large couverture du domaine (néanmoins un nombre de classes trop important peut suggérer une ontologie difficile à utiliser dans la pratique). En ce qui concerne le nombre total de classes, ONTOPARON se situe entre les deux autres ontologies. L'ontologie ONTOAD, pour la maladie d'alzheimer, est la plus grande, avec 5899 classes (*versus* les 2898 classes de ONTOPARON). Cela semble raisonnable, car en plus de modéliser les aspects sociaux / environnementaux et quotidiens de la maladie (comme ONTOPARON), elle modélise également la biologie de la maladie, par exemple des fonctions moléculaires (telles que l'activité tau-protéine kinase), qui est en dehors du domaine de ONTOPARON.

En revanche, ONTOPARON est beaucoup plus grande que l'Ontologie de coordination des soins infirmiers, avec 2989 classes contre 400 pour NCCO. Cependant, si l'on compare uniquement le module de coordination de ONTOPARON à celui de NCCO, on constate qu'ils sont beaucoup plus proches, à 250 ou 400 classes, ce qui semble raisonnable puisque NCCO contient d'autres classes en plus de celles liées à la coordination des soins infirmiers. Une analyse plus approfondie de cette différence sera poursuivie à l'aide d'outils d'analyse ontologique, en particulier OnAGUI [2], utilisés pour aligner les concepts d'ontologies.

Nombre de termes, français et anglais : ces quantités sont des estimations de l'état de préparation de l'ontologie à mettre en correspondance avec d'autres ontologies ; le cas d'utilisation nécessite des termes en français, mais la mise en correspondance avec d'autres ontologies nécessite les termes en anglais correspondants.

Profondeur médiane des classes : Il s'agit d'une estimation de haut niveau de la complexité des ontologies, avec une plus grande profondeur suggérant une plus grande complexité. Nous comparons les profondeurs par leurs médianes plutôt que par leurs moyennes car les profondeurs ne sont pas normalement distribuées - données non montrées. Comme le montre le tableau 3, les profondeurs médianes des classes du NCCO et de ONTOAD sont plutôt extrêmes, à 3 (NCCO) et 9 (ONTOAD). En revanche, la profondeur médiane des classes d'ONTOPARON est 4. Ceci est cohérent avec une hiérarchie d'héritage raisonnable.

FIGURE *3 – Comparaison qualitative de OntoParon avec NCCO et OntoAD.*

| | OntoParon | | | | | NCCO | OntoAD |
|---|---|---|---|---|---|---|---|
| **Ontologie** | **Noyau** | **Médical** | **Socio-environ.** | **Coordin.** | **Total** | **Total** | **Total** |
| **Classes** | 414 | 1313 | 921 | 250 | 2898 | 400 | 5899 |
| **Termes (fr)** | 433 | 1239 | 1033 | 282 | 2987 | 0 | 3283 |
| **Termes (en)** | 52 | 76 | 212 | 12 | 352 | 400 | 5899 |
| **Prof. moy.** | 3 | 5 | 4 | 4 | 4 | 3 | 9 |

## 4 Quelques bonnes pratiques

### 4.1 Sur la modularisation elle-même

La construction s'opère donc sur trois types de modules : (1) un module top-core, (2) des modules spécifiques et (3) un module de consolidation. Comme suggeré dans Aimé *et al.* (2016), chaque module répond à un certain nombre de contraintes parmi lesquelles nous pouvons citer :

— tout module de consolidation possède son propre espace de nommage ;

2. https://github.com/lmazuel/onagui

— tout module de consolidation répond à une problématique (i.e. un besoin de modélisation d'un point de vue donné) ;
— tout module de consolidation ne peut hériter que d'un ou de plusieurs modules spécifiques, ou (non exclusif) d'un ou de plusieurs modules de consolidation ;
— toute relation d'objet d'un module de consolidation a pour domaine et codomaine des concepts dans des modules différents (sinon elles doivent être créées au sein du module concerné).

Chaque module spécifique hérite directement de tous les concepts et relations de l'ontologie noyau, et indirectement de la top-ontologie si elle est définie. Chaque module spécifique exprime un point de vue ou un sous-domaine.

Un module de consolidation permet la spécialisation de concepts en provenance de modules spécifiques par l'ajout de labels, de relations avec des concepts ou d'attributs. Les modules de consolidation permettent de former des ensembles logiques et cohérents de modules spécifiques ou de consolidation en fonction de leur finalité d'usage. Ce jeu de modules consolidés peut ensuite être assemblé, permettant d'avoir une grande souplesse dans la diffusion de la base de connaissances (pas besoin d'exporter plus que de besoin). D'autre part, chaque modification dans un module spécifique (ou même de consolidation pour peu qu'il soit importé dans un autre module de consolidation) est automatiquement reportée dans l'ensemble des modules héritant directement ou indirectement (par transitivité).

### 4.2 La nécessité d'une ontologie noyau

Quand on passe de l'ontologie monolithique à l'ontologie modulaire, on s'aperçoit rapidement qu'une ontologie noyau est indispensable. Dans notre cas, nous avions, comme souvent dans le processus de construction d'ontologies de notre laboratoire, commencé par une ontologie non modulaire et sans ontologie noyau même si on savait qu'elle deviendrait assez rapidement nécessaire. En effet et par exemple, trouver des concepts qui subsument des actes de différents personnels médicaux ou de coordination crée le besoin d'avoir la notion d'acte non spécifique. Et ce type de constat se répétant n fois, l'ontologie noyau devenait indispensable. Nous avons à notre disposition, une ontologie noyau subsumée par une top que nous réutilisons habituellement, TOP-MENELAS et c'est celle-ci qui a été retenue pour ONTOPARON avec quelques aménagements [3].

### 4.3 Ne travailler qu'avec des ontologies

L'importation d'ontologies implique que la qualité de la modélisation dépend de la qualité des ontologies importées. Sans chercher à mesurer cette qualité de façon précise, il faut faire attention à des problèmes de modélisation dans les serveurs de terminologies comme bioportal [4]. Avec des arguments de cohérence de langage , des Systèmes d'Organisation des Connaissances (SOC) qui ne sont pas des ontologies, sont modélisées comme des ontologies avec, à la fin, des erreurs manifestes. Dans certains cas, les auteurs du SOC assument l'« erreur » comme pour le MeSH [5] mais ce n'est pas toujours le cas.

### 4.4 Mettre à disposition

La mise à disposition d'ontologies modulaires est plus compliquée qu'une ontologie monolithique car l'importation des ontologies standards disponibles sur des serveurs n'est pas toujours possible. Pour prendre 2 des ontologies les plus utilisées sur Internet, FOAF et Dublin Core (DC), au moment de l'écriture de ce papier, FOAF est accessible directement dans les interfaces du logiciel Protégé (et ça marche !) alors que DC n'est pas immédiatement accessible. Une solution est alors de regrouper dans un même dossier l'ensemble des modules

3. http://bioportal.bioontology.org/ontologies/TOP-MENELAS
4. http://bioportal.bioontology.org
5. http://bioportal.bioontology.org/ontologies/MESH

nécessaires à la reconstruction de l'ontologie, ceux qui forment le travail de l'équipe et ceux qui correspondent aux ontologies standards. En termes de mise à disposition vers l'extérieur, on a alors des stratégies différentes selon les cas : $a$) on veut mettre l'ensemble des fichiers à disposition aux utilisateurs comme les modules d'un programme et on les distribue via, par exemple, un Github ou $b$) on veut mettre le résultat final de l'ontologie à disposition, toute importation faite, et on utilise la fonction *Merge* de Protégé.

### 4.5 La modularité des méta-données

Des problèmes de chargement des ontologies (ontologies inaccessibles par importation directes, serveur amenant sur des pages complexes où trouver la bonne ontologie n'est pas évident) et des facilités des ontologistes (n'utilisant par exemple qu'une méta-donnée du *Dublin Core*) amènent les auteurs à ne vouloir renseigner que quelques méta-données intéressantes. Force est de constater à ce moment que les classes sont souvent, ou mal nommées (erreur dans l'URI) ou mal organisées. Ainsi quelqu'un se servant des `skos:prefLabel` et `skos:altLabel` peut vouloir les rentrer à la main dans l'ontologie et oublier qu'il faut les ranger sous `rdfs:Label` au risque d'égarer un logiciel qui fouillerait les termes d'une ontologie via une requête sur les seuls `rdfs:label`. La solution est dans $a$) la recherche et la récupération du bon fichier sur Internet pour le mettre dans son propre dossier de travail en acceptant souvent de récupérer des méta-données *a priori* inutiles (en suivant le même exemple, on peut ramener dans son dossier de travail pour importation à l'ouverture le fichier `https://www.w3.org/2009/08/skos-reference/skos.rdf`) ou $b$) la reconstruction d'un fichier spécifique associé à ceux de l'ontologie dans le même dossier et mis à disposition en même temps. Dans tous les cas, autant que faire se peut, on gère l'importation de ces méta-données via une seule ontologie, la noyau sus-nommée.

## 5 Perspectives et conclusion

Le travail décrit ici est en cours et ouvre quelques perspectives que nous abordons rapidement.

En termes de modularité de l'ontologie, notre prochaine étape est de travailler sur les relations et leurs positionnements sur les différents modules, en respectant les principes exprimés dans 4.1 mais qui nécessitent quelques approfondissements pour tenir compte de toutes les situations des domaine et codomaine des relations. C'est ce à quoi nous allons rapidement nous attacher en réimplémentant et réorganisant les relations de l'ontologie monolithique d'origine dans sa version modulaire.

En termes d'études de la modularité et de la caractérisation des modules les uns par rapport aux autre, Abbès *et al.* (2012) ont proposé un certain nombre de mesures que nous pensons implémenter pour mieux décrire les propriétés de nos modules.

Enfin, par rapport à l'ontologie décrite ici, notre objectif à long terme est d'appliquer ce modèle ontologique modulaire à d'autres pathologies neurologiques comme la maladie de Parkinson. La vue modulaire pourrait nous permettre de réimplémenter uniquement les modules spécifiques aux maladies tout en réutilisant les autres. la modularité peut aussi nous permettre de n'utiliser qu'un module spécifique pour observer uniquement la dimension socio-environnementale par exemple, ou encore de comparer des modèles de coordination qui peuvent être différents selon les structures et comprendre ainsi leur impact dans les parcours de santé des patients.

De façon plus générale, il est clair qu'une ontologie modulaire est la conception la plus adaptée à notre utilisation : comprendre à la fois les ruptures de parcours, comprendre les besoins des patients et des familles et modéliser les actions de coordination déployées – quelles sont les demandes faites aux coordinateurs, par qui et à quel moment du parcours, comprendre les solutions mises en place pour répondre aux demandes, quel type d'action est déployé : action de prévention, action de diffusion des informations, action de formation, de soutien, de recherche ... – au cours de cette coordination. Le fait que les modules de l'ontologie soient encore en cours de raffinement offre l'opportunité d'étudier l'évolution de l'ontologie dans

son ensemble et comment ces modules contribuent à sa gestalt. Cela est fait ici par la comparaison de l'ontologie modulaire en évolution avec des ontologies monolithiques qui sont pertinentes pour le domaine : une ontologie d'une maladie connexe et une ontologie de la fourniture de soins. Cette comparaison pose un certain nombre de questions de recherche ouvertes, telles que celle de la profondeur optimale d'une telle ontologie et de ses modules individuels ; le développement continu de ONTOPARON permettra de chercher les réponses.

## Références


ABBÈS S. B., SCHEUERMANN A., MEILENDER T. & D'AQUIN M. (2012). Characterizing modular ontologies. In *Proceedings of the 6th International Workshop on Modular Ontologies, Graz, Austria, July 24, 2012*.

AIMÉ X., GEORGE S. & HORNUNG J. (2016). VetiVoc : A modular ontology for the fashion, textile and clothing domain. **11**(1), 1–28.

BAO J. & HONAVAR V. (2006). Adapt OWL as a Modular Ontology Language.

CORCIA P., PRADAT P., SALACHAS F., BRUNETEAU G., LE FORESTIER N., SEILHEAN D., HAUW J. & MEININGER V. (2008). Causes of death in a post-mortem series of ALS patients. **9**(1), 59–62.

DRAMÉ K., DIALLO G., DELVA F., DARTIGUES J. F., MOUILLET E., SALAMON R. & MOUGIN F. (2014). Reuse of termino-ontological resources and text corpora for building a multilingual domain ontology : An application to Alzheimer's disease. *Journal of Biomedical Informatics*, **48**, 171–182.

FRIEDMAN C., BORLAWSKY T., SHAGINA L., XING H. R. & LUSSIER Y. A. (2006). Bio-ontology and text : bridging the modeling gap. **22**(19), 2421–2429.

GRAU B. C., PARSIA B., SIRIN E. & KALYANPUR A. (2006). Modularity and web ontologies. In *Proceedings of the 20$^{th}$ International Conference on Principles of Knowledge Representation and Reasoning*, p. 198–209 : AAAI Press.

LAVERNHE S., ANTOINE J.-C., COURT-FORTUNE I., DIMIER N., COSTES F., LACOUR A. & CAMDESSANCHÉ J.-P. (2017). Home care organization impacts patient management and survival in ALS. **18**(7), 562–568.

LIN J. (2008). Scalable language processing algorithms for the masses : A case study in computing word co-occurrence matrices with mapreduce. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, p. 419–428, Stroudsburg, PA, USA : Association for Computational Linguistics.

PATHAK J., SOLBRIG H. R., JOHNSON T. M., BUNTROCK J. D. & CHUTE C. G. (2009). Survey of modular ontology techniques and their applications in the biomedical domain. (3), 225–242.

POPEJOY L. L., KHALILIA M. A., POPESCU M., GALAMBOS C., LYONS V., RANTZ M., HICKS L. & STETZER F. (2014). Quantifying care coordination using natural language processing and domain-specific ontology. *Journal of the American Medical Informatics Association*, p. 1–9.

SORIANI M.-H. & DESNUELLE C. (2017). Care management in amyotrophic lateral sclerosis. **173**(5), 288–299.

SPASIC I., ANANIADOU S., MCNAUGHT J. & KUMAR A. (2005). Text mining and ontologies in biomedicine : making sense of raw text. **6**(3), 239–251.

# Une ontologie pour la formalisation et la visualisation des connaissances scientifiques

Vincenzo Daponte[1], Gilles Falquet[1]

[1] CUI, Centre Universitaire d'informatique, Université de Genève, Genève, Suisse
Vincenzo.Daponte@unige.ch, Gilles.Falquet@unige.ch

**Résumé** : La construction d'une ontologie des objets de connaissance scientifique, présenté ici, s'inscrit dans le développement d'une approche orientée à la visualisation des connaissances scientifiques. Il est motivé par le fait que les concepts d'organisation de la connaissance scientifique (théorème, loi, expérience, preuve, ...) apparaissent dans des ontologies existantes mais qu'aucune de celles-ci n'est centrée sur cette thématique et n'en présente une organisation simple et facilement utilisable. Nous présentons la première version construite à partir de sources ontologiques (ontologies des objets de connaissance de certains domaines, lexicales et de niveau supérieur), de bases de connaissances spécialisées et d'interviews avec des scientifiques. Nous avons aligné cette ontologie avec certaines des sources utilisées, ce qui a permis de vérifier sa consistance par rapport à ces dernières. La validation de l'ontologie consiste à l'utiliser pour formaliser des connaissances de diverses sources, ce que nous avons commencé à faire dans le domaine de la physique.

**Mots-clés** : Ontologies, Connaissance scientifique, Visualisation des connaissance

## 1 Motivations

L'accès à la connaissance scientifique, qu'elle soit générale ou factuelle doit forcément passer par une présentation visuelle, auditive, ou autre qui fait appel à un ou plusieurs sens de l'être humain. Si on s'intéresse à la présentation visuelle des connaissances, on constate que la langue naturelle écrite y occupe une place prépondérante mais que d'autres formes graphiques (notations, formules mathématiques et chimiques, diagrammes, tableaux, formulaires, hypertextes , etc.) y jouent un rôle important pour faciliter l'accomplissement de diverses tâches intellectuelles (calcul, comparaison, déduction, etc.).

Le cadre général dans lequel s'inscrit notre travail est l'étude des techniques de visualisation des connaissances scientifiques et en particulier leur spécification formelle en vue des construire des outils de visualisation adaptés aux tâches de l'utilisateur scientifique. En effet, l'expérience montre qu'il n'existe pas *une* technique optimale de visualisation mais que l'efficacité de chaque technique dépend du contexte et des objectifs de l'utilisateur (voir, par exemple Card *et al.* (1999)).

Pour représenter formellement la notion de technique de visualisation il faut, suivant le model de référence proposé par Chi (2000), définir un modèle abstrait des données à visualiser, un modèle abstrait des objets visuels et une application du modèle de données dans le modèle visuel abstrait. Dans le cas de la visualisation de connaissances scientifiques il faut donc créer un modèle abstrait des connaissances scientifiques à visualiser. Notre but étant de fournir une formalisation de la visualisation qui s'applique à n'importe quelle science, nous avons décidé de construire une ontologie des objets de structuration de la connaissance utilisés dans les diverses sciences. Une telle ontologie rendra possible

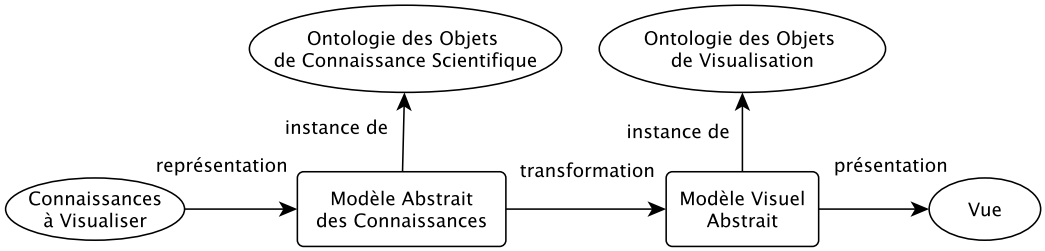



FIGURE *1 – Adaptation du modèle de Chi (2000) à la visualisation des connaissances scientifiques*

la modélisation des connaissances à visualiser sous forme d'instances des classes de cette ontologie, selon le schéma de la figure 1.

Si les sciences varient en fonction de leurs objets d'étude (les fonctions différentiables, les papillons, les sociétés humains, les particules élémentaires, ...), elles possèdent également leurs propres concepts pour structurer la connaissance produite. Les mathématiques produisent des théorèmes, corollaires, lemmes, conjectures, preuves alors que la physique parle plutôt de lois, de principes, de mesures, de résultats expérimentaux où que l'anthropologie produit des observations, théories, hypothèses explicatives, méthodes. De plus, chaque science a développé un ensemble de formalismes (donc des constructions linguistiques) pour exprimer et traiter ces connaissances : formules et équations chimiques, formules mathématiques, diagrammes de flux, diagrammes d'interaction, arbres syntaxiques, etc. et un ensemble de techniques de productions de la connaissance : expérimentation, raisonnement formel, enquêtes, observations, etc. On constate cependant que le vocabulaire de classification des objets de connaissance scientifiques est restreint et que plusieurs termes recouvrent souvent des concepts similaires. Par conséquent on peut penser arriver à construire une ontologie centrale composée d'un petit nombre de classes d'objets scientifiques. Les connaissances à visualiser pourront alors être représentées comme des instances de ces classes.

Dans le reste de cet article nous commencerons par examiner les travaux centrés sur la représentation des connaissances scientifiques et ceux qui, tout en étant plutôt connexes, peuvent fournir des éléments importantes. Nous présenterons ensuite la méthode suivie pour créer une première version de l'ontologie SKOO des objets de connaissance scientifique et l'ontologie obtenue. Nous présenterons ensuite les premières évaluations que nous avons réalisées. Dans la conclusion nous donnerons des perspectives sur l'utilisation pratique de cette ontologie et sur la poursuite des travaux d'évaluation et de mise au point de l'ontologie.

## 2 Etat de l'art

À notre connaissance il n'existe pas à l'heure actuelle d'ontologie dont le champ est la conceptualisation des objets servant à représenter ou structure la connaissance scientifique en général. Notons aussi que les travaux d'épistémologie des sciences ne s'attaquent en général pas à cette question ontologique générale mais traitent soit d'une science parti-

culière, soit d'un aspect particulier des sciences. Ce travail ontologique a par contre été réalisé en ingénierie des connaissances pour quelques domaines spécifiques.

Le noyau de l'ontologie OMDoc présenté dans (Lange (2013)) présente une modélisation de la connaissance mathématique sous forme d'éléments de connaissances (*knowledge items*) qui sont des types d'objets mathématiques, des théories ou des énoncés (*statements*), lesquels peuvent être de type *Assertion*, *Proof*, *Definition*, *Axion*, etc. De plus, des relations lient ces types d'éléments entre eux, par exemple la relation *proves* lie *Proof* à *Assertion*. Bien que cette ontologie soit dédiée aux mathématiques, on voit qu'elle pourrait s'étendre facilement à d'autres sciences.

L'ontologie SIO (Semanticscience Integrated Ontology), Dumontier *et al.* (2014), est une ontologie de niveau supérieur qui vise essentiellement la représentation des connaissances biomédicales. Le but premier de SIO est la descriptions des objets complexes du domaine biomédical (avec des relations composant-composé) et des processus dans lesquels ils interviennent ou des procédures (expérimentales) qu'on leur applique. On trouve cependant dans la classe *description* de SIO bon nombre de concepts servant à structurer la connaissance scientifique : *argument*, *belief*, *conclusion*, *evidence*, *hypothesis*, ... Mais contrairement à OMDoc, il n'y a cependant pas de relations spécifiques entre ces concepts, ils peuvent par contre être liés aux objets qu'ils décrivent. SIO décrit également les objets linguistiques, mathématiques et les médias servant de support à la connaissance.

La notion de processus scientifique expérimental est au centre du système EXPO (Soldatova & King (2006)). Il associe l'ontologie SUMO ((Niles & Pease, 2001, ))) à des ontologies d'expériences spécifiques à un sujet en formalisant les concepts génériques de conception expérimentale, de méthodologie et de représentation des résultats. EXPO vise à décrire différents domaines expérimentaux et à donner une description formelle des expériences pour l'analyse, l'annotation et le partage des résultats.

On peut également considérer ce qui a été fait dans les bases de connaissances scientifiques, telles Gene Ontology (Ashburner *et al.* (2000)), OntoMathPro (Nevzorova *et al.* (2014)) ou encore FMA (Rosse & Mejino (2007)) qui ont pour but de représenter l'état actuel de nos connaissances dans un domaine. Elles sont en général constituées d'une partie terminologique qui organise très précisément les concepts du domaine et d'une partie composée d'assertions (énoncés) qui représentent nos connaissances à propos de ces concepts. Dans Gene Ontology (GO) les énoncés sont appelés « annotations ». Ils lient typiquement un gène et un terme de l'ontologie GO (par exemple pour indiquer que le gène possède une certaine fonction). Les énoncés sont qualifiés par un type de preuve (expérimentale, inférence phytogénétique, inférence automatique, ...). Dans OntoMathPro (Nevzorova *et al.* (2014))les niveaux terminologiques et assertions existent bel et bien mais ne sont pas structurellement séparés. Ainsi le théorème de Stokes (assertion) n'est pas une instance mais une sous-classe de la classe *Theorem*. De même, on trouve comme sous-classe de premier niveau de *Mathematical knowledge object* aussi bien *Theorem* que *Tensor*. Il y a donc agrégation des objets de description de la connaissance et des objets du domaine sur lequel on travaille.

Notons encore qu'il existe des ontologies dont le but est uniquement de répertorier et classer les objets d'étude d'un domaine ou de créer une terminologie d'une domaine (SWEET, ScienceWISE, ...). En général ces ontologies ne s'intéressent pas aux objets de structuration de la connaissance.

À l'inverse, on trouve dans une ontologie lexicale comme WordNet un grand nombre de concepts tels que théorème, loi, définition, hypothèse, corollaire. Cependant on remarque

que ces concepts ne sont pas organisés de manière directement utilisable. On a par exemple les chaines de relation d'hyperonymie

$$theorem < idea < content < cognition$$

et

$$corollary < ... < process < content < cognition$$

alors que d'un point de vue formel un corollaire est un théorème. En d'autres termes, on ne peut pas extraire une ontologie des objets de connaissance scientifique par simple projection d'une partie de WordNet. Il en va de même pour d'autres ontologies de niveau supérieur (SUMO, CyC, ...)

## 3 Construction de l'ontologie SKOO

Pour construire l'ontologie des objets de connaissance scientifique (*Scientific Knowledge Objects Ontology* - SKOO) nous avons appliqué le processus suivant :

1. Nous avons collecté un ensemble de termes utilisés pour structurer les connaissances dans différents domaines scientifiques. Cette opération a été effectuée par consultation d'ouvrages (manuels, formulaires, monographies « handbooks ») en biochimie, physique, mathématiques, linguistique, sociologie ; interviews avec des scientifiques de différents domaines ; analyse du niveau terminologique de bases de connaissances et d'ontologies scientifiques (Gene Ontology, OntoMathPro, ...)
2. Pour construire le niveau supérieur de l'ontologie nous avons tout d'abord associé les termes mis en évidence à des « synsets » équivalents ou plus généraux de WordNet. Puis nous avons utilisé l'ontologie DOLCE, déjà alignée avec WordNet, pour trouver des concepts de niveau supérieur.
3. Enfin nous avons défini des relations entre les concepts du niveau supérieur à partir de relations trouvées dans des ontologies scientifiques, en particulier OMDoc, et par spécialisation de certains relations de haut niveau de DOLCE.

La figure 2 montre le niveau supérieur de l'ontologie obtenue et ses liens avec DOLCE et WordNet. Nous décrivons ci-dessous l'interprétation de chacune de ses classes.

**Sci_Knowledge_Item** Les éléments de connaissance scientifique sont tous les objets qui servent à structurer l'expression de la connaissance scientifique. Il peut s'agir d'objets, tels les théorèmes, lois (physique, chimiques), modèles ou méthodes qui portent en eux-mêmes de la connaissance au sens platonicien de croyance vraie et justifiée. Mais il peut aussi s'agir d'objets « auxiliaires » tels que les définitions, exemples, preuves, hypothèses, problèmes. Ces objets correspondent aux objets de la classe *description* de l'ontologie DOLCE (Masolo *et al.* (2003)).

**Sci_Information_Object** Cette classe a pour but de regrouper toutes les formes d'expression des éléments de connaissance, qu'elles soit linguistique ou sous forme de diagramme, schémas, formules, etc.. Il s'agit d'une sous-classe de la classe

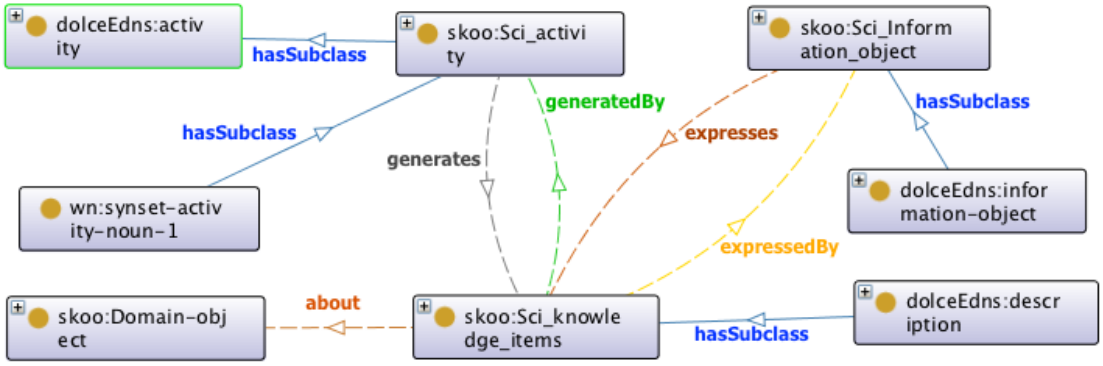



FIGURE 2 – *Structure du niveau supérieur de l'ontologie SKOO*

*information-object* de l'ontologie DOLCE (Masolo *et al.* (2003)), et sa classe principale *Sci-linguistic-object* est une sous-classe de la classe *linguistic-object* de DOLCE (Masolo *et al.* (2003)). Cette classe veut inclure toutes les formes et méthodes pour exprimer les concepts utilisés pour représenter la connaissance des disciplines considérées.

**Sci_Activity** Cette classe représente les activités, au sens de *activity* (hyponyme de *human activity*) dans WordNet, qui servent à engendrer des éléments de connaissances scientifique. Il peut s'agir d'activités de type expérimental (processus, expérimentation, observation), mais aussi de type empirique (mener des enquêtes) ou formelles (prouver formellement, calculer). La description précise des activités, en particulier des expérimentation, n'est pas définie dans cette ontologie car elle est déjà couverte par d'autres ontologies, telles SIO et EXPO.

***Domain-object*** représente tous les objets à propos desquels des connaissances scientifiques sont exprimées. Cette classe sert de point d'ancrage aux classes décrivant les objets étudiés dans des domaines spécifiques. Lors de l'usage pratique de l'ontologie, le principe est d'importer une ontologie d'objets du domaine scientifique concerné et créer des axiomes de subsomption $C \sqsubseteq Domain - object$ pour ses classes de niveau supérieur.

## 4 Evaluation

Nous avons mené deux types d'évaluation de type consistance et capacité. En plus de la consistance interne de l'ontologie, pour donner une indication de la consistance externe (par rapport à d'autre ontologies), l'ontologie a été alignée avec les ontologies OMDoc, DOLCE et WordNet. Pour cela nous avons traduit sous forme de classes OWL les concepts de OMDoc et WordNet, puis nous avons créé des axiomes de correspondance de type owl :subClassOf et owl :EquivalentClass entre celles-ci et SKOO. La table 1 montre quelques uns de ces axiomes. Nous avons ensuite vérifié la consistance de l'ontologies

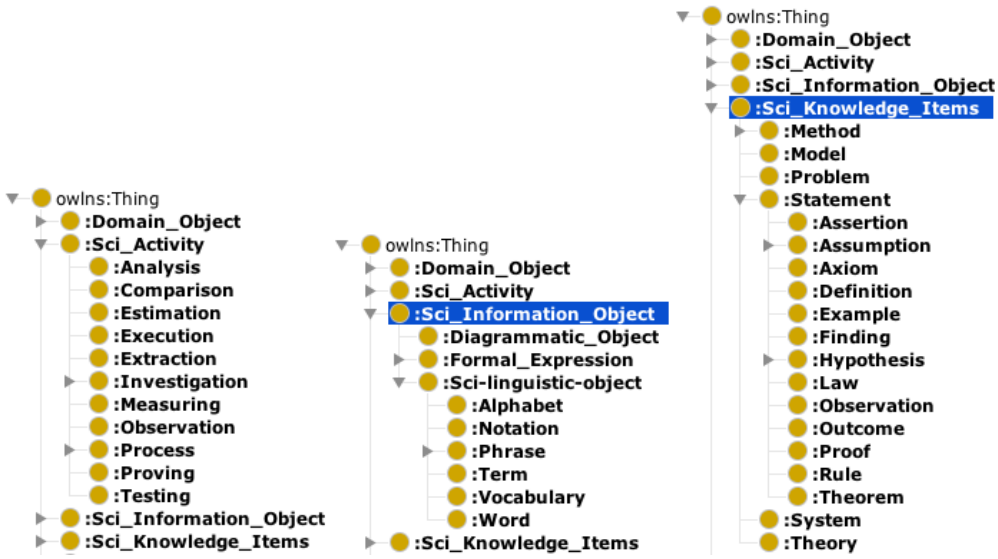



FIGURE *3 – Les niveau supérieurs des sous-classes de* Sci-activity, Sci-knowledge-items *et* Sci-information-object.

| SKOO | OMDoc | DOLCE | WordNet |
|---|---|---|---|
| Sci-knowledge-items | $\sqsupseteq$ MathKnowledgeItem | $\sqsubseteq$ description | |
| Statement | $\sqsupseteq$ Statement | | $\sqsubseteq$ statement |
| Theory | $\sqsupseteq$ Theory | $\sqsubseteq$ theory | $\sqsubseteq$ theory |
| Assertion | $\sqsupseteq$ Assertion | | $\sqsubseteq$assertion |
| Axiom | $\sqsupseteq$ Axiom | | $\sqsubseteq$axiom |
| Definition | $\sqsupseteq$ Definition | | $\sqsubseteq$definition |
| Proof | $\sqsupseteq$ Proof | | $\sqsubseteq$ proof |
| Sci-activity | | | $\sqsubseteq$ activity |
| Process | | $\sqsupseteq$ activity | $\sqsubseteq$ process |
| Sci-information-object | | $\sqsubseteq$ information-object | |

TABLE *1 – Correspondances entre les classes SKOO et OMDoc, DOLCE, WordNet*

obtenue en fusionnant SKOO, les trois ontologies et les axiomes de correspondance (mais sans correspondances entre DOLCE, WordNet et OMDoc).

Pour évaluer les capacités de l'ontologie nous devons vérifier si, étant donné un système de visualisation de connaissances scientifiques, l'ontologie permet de créer un modèle abstrait adéquat pour ces connaissances. Dans le cas des bases de connaissances structurées et homogènes, par exemple les annotations de Gene Ontology ou des formulaires mathématiques il est aisé de vérifier l'adéquation de l'ontologie. En effet ces connaissances correspondent généralement à des énoncés (*Statement*) qui peuvent être des théorèmes. Par contre le cas des connaissances exprimées dans des textes est plus complexe. Nous avons effectué un premier test en prenant comme système de visualisation une partie d'un

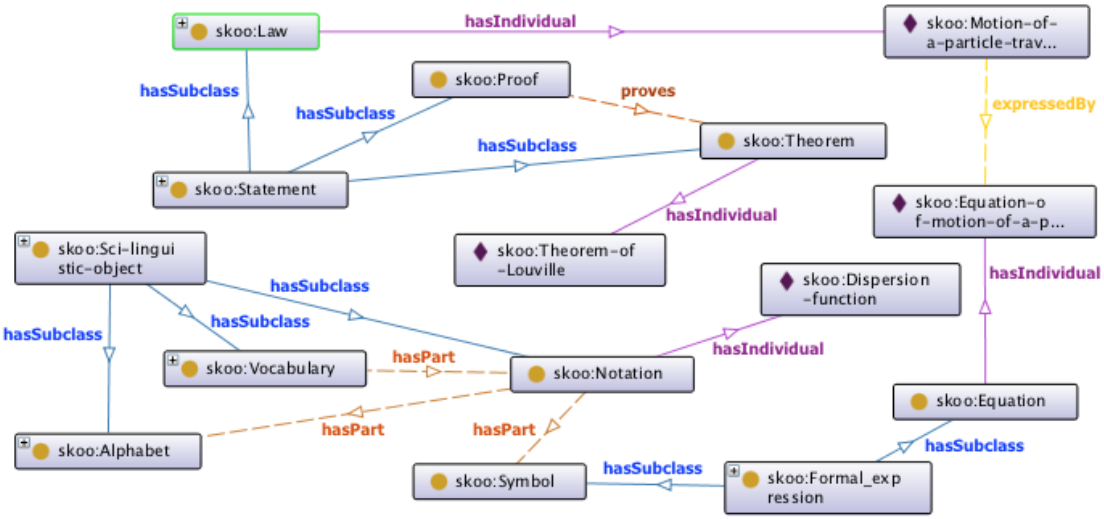



FIGURE 4 – *Quelques concepts contenus dans plusieurs sections du chapitre 3 de (Wille (2000)) exprimés sous forme d'instances de l'ontologie SKOO.*

ouvrage de physique des accélérateurs (Wille (2000)). Nous avons modélisé différents concepts des différentes sections du chapitre 3. En particulier, le concept principal exprimé dans la section 3.2, la section 3.6 complète de cet ouvrage (*dispersion and momentum compaction factor*) et un théorème utilisé dans la section 3.8 sous forme d'instances de classes SKOO. La figure 4 montre la modélisation d'une loi physique (instance de *Law*) représentée par une équation (instance de *Equation*) et aussi le modélisation d'un théorème (instance de *Theorem*) et d'une notation (instance de *Notation*) utilisée pour représenter un concept particulier. Il faut souligner que dans les relations montrées par cette figure, la relation *hasIndividual* est utilisée par *Protégé* pour associer le type de l'instance (la classe) à l'instance elle-même.

## 5 Conclusions et travaux futurs

Nous avons présenté la construction de la première version de l'ontologie SKOO dont le but est de fournir un modèle général pour la modélisation de connaissances scientifiques à visualiser (selon le schéma de Chi (2000)). Bien que les concepts représentés dans cette ontologie existent tous dans d'autres ontologies, aucune de celles-ci ne les regroupe de manière à être directement utilisables pour représenter la connaissance scientifique. D'où l'intérêt de l'ontologie SKOO. L'ontologie a été alignée avec des ontologies de référence pour vérifier à l'aide d'un raisonneur qu'elle n'entraient pas en contradictions avec celles-ci. D'autre part nous avons commencé la validation de la capacité de cette ontologie à modéliser les connaissances représentées dans des bases de connaissances existantes, ce qui ne pose pas de problème particulier, et les connaissances représentées dans des (hyper)textes, ce qui est plus difficile, surtout pour les textes de sciences humaines et sociales. Après avoir réalisé un test sur une partie d'un ouvrage de physique, nous allons entre-

prendre des tests sur des ouvrages de sciences humaines et sociales.

La prochaine étape de ce travail consistera à modéliser complètement divers systèmes de visualisation de connaissances existants. Nous utiliserons pour cela le langage SPARQL pour spécifier les transformations d'un modèle de connaissance (exprimé avec SKOO) vers un modèle d'objets de visualisation (comprenant des listes, arbres, graphes, textes, formes géométriques, etc.). Ceci permettra de valider le modèle complet de spécification de techniques de visualisation. À partir de là il sera possible de créer un système de génération de visualisations à partir de leur spécification et de l'utiliser pour créer de nouvelles techniques de visualisation et de les tester avec des utilisateurs.

Au cours de ce travail nous nous sommes aperçus que l'intérêt de cette ontologie va au-delà de la seule visualisation des connaissances. Elle est par exemple applicable dans le cadre de la recherche d'information précise ou du raisonnement automatique sur de grands ensembles de connaissances scientifiques.

*Remerciements*

Ce travail a été enrichi par les précieux conseils et l'expérience de Giuseppe Cosenza et Jean-Pierre Hurni, que les auteurs souhaitent remercier.

## Références

ASHBURNER M., BALL C. A., BLAKE J. A., BOTSTEIN D., BUTLER H., CHERRY J. M., DAVIS A. P., DOLINSKI K., DWIGHT S. S., EPPIG J. T. *et al.* (2000). Gene ontology : tool for the unification of biology. *Nature genetics*, **25**(1), 25.

CARD S. K., MACKINLAY J. D. & SCHNEIDERMAN B. (1999). *Readings in Information Visualization : Using Vision to Think*. Morgan Kaufman.

CHI E. (2000). A taxonomy of visualization techniques using the data statereference model. In *InfoVis 2000. IEEE Symposium*, p. 69–75.

DUMONTIER M., BAKER C. J., BARAN J., CALLAHAN A., CHEPELEV L., CRUZ-TOLEDO J., DEL RIO N. R., DUCK G., FURLONG L. I., KEATH N., KLASSEN D., MCCUSKER J. P., QUERALT-ROSINACH N., SAMWALD M., VILLANUEVA-ROSALES N., WILKINSON M. D. & HOEHNDORF R. (2014). The semanticscience integrated ontology (sio) for biomedical research and knowledge discovery. *Journal of Biomedical Semantics*, **5**(1), 14.

LANGE C. (2013). Ontologies and languages for representing mathematical knowledge on the semantic web. *Semantic Web*, **4**(2), 119–158.

MASOLO C., BORGO S., GANGEMI A., GUARINO N. & OLTRAMARI A. (2003). *The WonderWeb library of foundational ontologies and the DOLCE ontology. WonderWeb (EU IST project 2001-33052) deliverable D18*. Rapport interne, LOA-ISTC-CNR.

NEVZOROVA O. A., ZHILTSOV N., KIRILLOVICH A. & LIPACHEV E. (2014). Ontomath pro ontology : a linked data hub for mathematics. In *International Conference on Knowledge Engineering and the Semantic Web*, p. 105–119 : Springer.

NILES I. & PEASE A. (2001). Towards a standard upper ontology. In *Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001*, p. 2–9 : ACM.

ROSSE C. & MEJINO J. (2007). The foundational model of anatomy ontology. In *Anatomy Ontologies for Bioinformatics : Principles and Practice*, volume 6, p. 59–117 : Springer.

SOLDATOVA L. N. & KING R. D. (2006). An ontology of scientific experiments. *Journal of the Royal Society Interface*, **3**(11), 795–803.

WILLE K. (2000). *The physics of particle accelerators : an introduction*. Clarendon Press.

# Construction d'ontologie pour le domaine du *sourcing*

Molka Tounsi Dhouib[1,2], Catherine Faron Zucker[1], Andrea Tettamanzi[1]

[1] UNIVERSITÉ CÔTE D'AZUR, INRIA, CNRS, I3S, Sophia Antipolis, France
{dhouib,faron,Tettamanzi}@i3s.unice.fr

[2] SILEX, FRANCE
{molka.tounsi}@silex-france.com

**Résumé** : Ces dernières années, de nombreuses entreprises s'orientent vers l'intégration du développement d'ontologies au sein de leurs processus pour mieux organiser les connaissances mises en jeu et améliorer les performances de leur traitement automatique. Dans cet article, nous décrivons notre travail de modélisation d'ontologie dans le domaine du *sourcing*, dans le but de décrire le contenu sémantique des offres et demandes de prestations. Nous nous focalisons sur la représentation des compétences et des domaines de compétences, dans le but de raisonner sur ces connaissances pour améliorer la recommandation des prestataires. Notre stratégie de développement d'ontologie repose sur (i) la réutilisation de référentiels existants pour représenter les compétences et domaines de compétences, (ii) la construction de référentiels internes et (iii) un travail d'alignement d'ontologies.

**Mots-clés** : Ingénierie des connaissances, construction d'ontologie, alignement d'ontologie, *sourcing*, plateforme B2B.

## 1 Introduction

Avec l'importance et la croissance exponentielle du volume des données des entreprises ou du Web, la disponibilité des ontologies (Gruber, 1993; Tchouanto Poosia, 2014; Uschold & King, 1995; Noy *et al.*, 2001) au sein des applications est devenue cruciale, car le traitement manuel de ces données devient impossible. L'exploitation d'ontologies ne se limite plus à des projets de recherche en intelligence artificielle, mais devient progressivement une réalité dans les projets industriels. De nos jours, plusieurs entreprises exploitent des ontologies comme élément clé de leur solution.

Le travail présenté dans cet article s'inscrit dans le cadre d'une collaboration entre la start-up Silex [1] et le laboratoire de recherche I3S. Silex est une plateforme de sourçage (*sourcing*) qui permet aux entreprises d'identifier les prestataires les plus adaptés à leurs projets. Nous commençons par définir le *sourcing* comme la démarche de recherche de fournisseurs répondant au mieux aux besoins, que ce soit en terme de coûts, de délais, d'innovation et de qualité. Cette démarche consiste à évaluer les fournisseurs selon plusieurs critères tels que la proximité, la capacité à répondre aux besoins, en termes de savoir-faire et capacité de production. Les méthodes traditionnelles de *sourcing* se basent sur l'utilisation de bases de données payantes en ligne et de bases spécialisées à un domaine précis, et sur les salons, fédérations et syndicats professionnels. Une nouvelle méthode de recherche des fournisseurs fait appel à des plateformes de e-sourcing (ESCHENLAUER, 2013) comme Silex.

Silex simplifie le processus du *sourcing* en permettant aux entreprises de publier leurs besoins en langage naturel. L'introduction d'ontologies dans le processus de recommandation de prestataires est une priorité pour Silex afin de conceptualiser ses connaissances et pouvoir normaliser et automatiser ce processus. Nous décrivons dans cet article notre travail d'ingénierie d'ontologie pour le domaine du *sourcing*. L'ontologie développée doit permettre de décrire le contenu sémantique des offres et demandes de prestation afin de raisonner sur ces connaissances et d'améliorer la recommandation des prestataires. Les aspects de partage et de réutilisation des ontologies sont parmi les raisons du succès de celles-ci. Mais les ontologies

1. https ://www.silex-france.com/silex/

reflètent en général le point de vue de leurs concepteurs. Le premier défi de notre travail est donc de définir une ontologie précise et complète pour représenter le domaine du *sourcing*, et qui soit en phase avec le point de vue de Silex. Dans ce contexte, nous avons identifié trois questions centrales dans notre travail d'ingénierie d'ontologie :

— Quels types de connaissances devons nous représenter afin de raisonner sur leur représentation et d'améliorer la qualité des recommandations des prestataires ?
— Quelles ontologies existantes pouvons nous réutiliser ?
— L'approche adoptée dans Silex nécessite-t-elle de construire une nouvelle ontologie ?

Dans la suite de l'article, la section 2 décrit avec plus de détails le contexte de ce travail. Nous détaillons ensuite notre méthode de construction d'ontologie pour le domaine du *sourcing* dans les sections suivantes. Enfin, nous concluons dans la section 6 en rappelant notre contribution et en évoquant nos perspectives.

## 2 Contexte

Dans le but de renforcer la qualité de la recommandation de prestataires à partir d'expressions de besoins en langue naturelle, Silex souhaite intégrer une couche sémantique au sein de sa plateforme B2B afin de permettre le traitement automatique des représentations des entreprises et de leurs demandes ou offres. Silex propose aujourd'hui aux entreprises partenaires de s'inscrire dans la plateforme en tant que donneurs d'ordre ou prestataires. Lors de leur inscription, les entreprises fournissent une description textuelle de leurs activités professionnelles, l'ensemble des offres qu'elles proposent et/ou des services qu'elles recherchent. L'objectif de Silex est d'analyser automatiquement ces descriptions textuelles afin d'établir un classement de prestataires les plus adaptés à un besoin donné.

Pour ce faire, des ontologies doivent être introduites au coeur du processus de recommandation afin de représenter le contenu des descriptions et d'élaborer des liens vers d'autres concepts qui peuvent être utilisés dans le même contexte et donc enrichir le processus de recommandation. Dans une première étape, nous avons convenu de nous limiter au domaine de l'informatique, qui est le principal domaine présent dans la plateforme. Le but à terme est cependant de couvrir tous les domaines professionnels des entreprises.

En étroite collaboration avec les responsables de Silex, et suite à une analyse détaillée des descriptions textuelles des besoins, offres et descriptions des entreprises, nous avons décidé de représenter trois types de connaissances : (i) la compétence peut être définie comme un ensemble de connaissances et de capacités requises dans la réalisation des tâches quotidiennes dans un domaine défini (Amourache *et al.*, 2008), (ii) la profession [2] désigne le métier exercé par une personne appartenant à un secteur d'activité particulier, (iii) le domaine d'activité [3] est le regroupement des entreprises de fabrication, d'industrie, de commerces ou de services qui ont la même activité principale.

Dans ce travail, nous adaptons la méthodologie NeON (Suárez-Figueroa & Gómez-Pérez, 2009) pour la construction de notre ontologie en nous basant sur le scénario de réutilisation, fusion et ré-ingénierie des ressources. Notre démarche se compose alors de : (i) la recherche des référentiels existants pour représenter les compétences, professions et domaines d'activités, (ii) la construction de référentiels internes à Silex pour représenter ses propres connaissances, (iii) l'alignement des ontologies.

Afin de déterminer la spécificité de notre ontologie, et avec l'aide des responsables de Silex, nous avons élaboré un ensemble de questions de compétences (Uschold & Gruninger, 1996) (Noy & McGuinness, 2000) :

Q1 : Quels compétences / métiers / domaines d'activités apparaissent dans une offre ou un besoin ?
Q2 : Quels sont les liens de parenté entre deux compétences / métiers donnés ?
Q3 : Quels sont les prestataires ayant un métier / compétence / domaine d'activité donné ?
Q4 : Quelles sont les compétences associées à un métier / domaine d'activité donné ?

2. https ://fr.wikipedia.org/wiki/Profession
3. https ://fr.wikipedia.org/wiki/Secteur_d%27activit%C3%A9

## 3 Identification et réutilisation de référentiels existants pour représenter les compétences, professions et domaines d'activités

Nous avons fait un état de l'art des référentiels existants et avons identifié les plus intéressants dans notre contexte. Nos critères de choix sont (i) la source du référentiel, (ii) sa mise à jour, (iii) les langues utilisées et (iv) son format.

### 3.1 Ontologie des compétences et des professions

#### 3.1.1 Classification des aptitudes, compétences, certifications et professions (ESCO)

ESCO [4] est une classification multilingue des aptitudes, compétences, certifications et professions européennes permettant de recenser et catégoriser ces derniers. L'objectif d'ESCO est de fournir une terminologie de référence commune pour améliorer le fonctionnement du marché du travail et combler les lacunes de communication entre les différents pays. ESCO est publiée au format SKOS-RDF et disponible dans les 24 langues officielles de l'UE et en islandais, norvégien, arabe et « anglais US ». ESCO est publiée en tant que LOD (Linked Open Data), ce qui permet de la connecter facilement à d'autres sources de données.

ESCO est composée de 5 modules inter-reliés : (i) professions, (ii) connaissances, (iii) aptitudes, (iv) compétences et qualifications, et (v) hiérarchie de la classification internationale des professions. Dans ce classement, nous n'utilisons que les professions et les compétences. ESCO contient 5 380 professions et 5 737 compétences. Dans cet article, nous nous intéressons à la modélisation du domaine informatique. En analysant manuellement l'arborescence d'ESCO, nous avons identifié que les métiers qui sont associés à ce domaine dérivent des concepts (i) « Professions intellectuelles et scientifiques », (ii) « Directeurs, cadres de direction et gérants », et (iii) « Professions intermédiaires ». Afin de représenter les compétences liées au domaine informatique, nous avons identifié le concept « Technologies de l'information et de la communication » qui contient les sous concepts suivants : (i) « Traitement de données numériques », (ii) « Création de contenus numériques », (iii) « Communication et collaboration numérique », (iv) « Sécurité des TIC » , et (v) « Résolution des problèmes à l'aide d'outils et de matériel TIC ». Au final, nous avons sélectionné manuellement 167 concepts qui représentent des métiers du domaine informatique et 524 concepts qui représentent les compétences liées à ce domaine.

#### 3.1.2 Répertoire Opérationnel des Métiers et des Emplois (ROME)

ROME [5] a été créé en 1989 en France par l'Agence Nationale pour l'Emploi (aujourd'hui Pôle Emploi). La version 2009 du ROME répertorie 531 fiches regroupant plus de 10 000 appellations différentes de métiers. ROME est publié sous forme de fichier Excel. Nous avons construit un thésaurus au format SKOS-RDF à partir de ce fichier. Le fichier Excel est composé de six colonnes. Les trois premières contiennent des informations de la version 2 de ROME : code, intitulé et appellation. Les trois contiennent les informations de même nature dans la version 3 de ROME. Dans notre thésaurus SKOS, nous avons décidé de garder ce lien entre les deux versions de ROME et cela par l'indication des anciennes appellations en valeurs de la propriété skos :altLabel.

Nous avons analysé les données de ROME afin de chercher les domaines et les métiers qui représentent le domaine de l'informatique. Nous avons sélectionné les sept domaines suivants : (i) Production et exploitation de systèmes d'information , (ii) Études et développement de réseaux de télécoms, (iii) Études et développement informatique, (iv) Expertise et support technique en systèmes d'information, (v) Administration de systèmes d'information, (vi) Conseil et maîtrise d'ouvrage en systèmes d'information et (vii) Direction des systèmes d'information. Au final, le thésaurus ROME que nous avons construit pour le domaine de l'informatique contient 118 concepts et 356 labels.

4. https ://ec.europa.eu/esco/portal/home

5. http ://www.pole-emploi.org/accueil/mot-cle.html ?tagId=94b2eaf6-d7bd-4244-bddc-01415605563b

### 3.1.3 Nomenclature des métiers des SI dans les grandes entreprises

Initiée par le Club Informatique des Grandes Entreprise Française (CIGREF) [6], cette nomenclature rassemble les descriptions des métiers présents dans les Directions des systèmes d'information. Elle contient 7 sous domaines de l'informatique et 36 noms de professions. Nous avons construit un thésaurus pour ce référentiel, qui contient 42 concepts de la même manière que pour ROME.

## 3.2 Ontologie des secteurs d'activité

Afin de représenter les domaines d'activité des entreprises, nous avons initialement identifié le référentiel NAF. Suite à plusieurs échanges avec les responsables de Silex, nous avons identifié la classification du site Kompass comme un autre bon candidat. Nous avons ensuite comparé ces deux référentiels d'un point de vue modélisation : (i) NAF présente sa classification en se basant seulement sur l'activité principale de l'entreprise alors que (i) Kompass établit sa classification en se basant sur les produits et les services des entreprises. De ce fait, le dernier niveau de la classification de NAF comprend seulement 732 activités alors que Kompass recense 55450 produits et services. Au final, Kompass se présente comme une classification plus exhaustive car elle permet de classer les activités d'une entreprise sous plusieurs catégories.

### 3.2.1 Nomenclature d'activités française (NAF)

NAF [7] est une nomenclature des activités économiques productives, principalement élaborée pour faciliter l'organisation de l'information économique et sociale. Afin de faciliter les comparaisons internationales, NAF est structurée de la même façon que la nomenclature d'activités européenne NACE, elle-même dérivée de la nomenclature internationale CITI. Nous avons choisi d'utiliser NAF au lieu d'utiliser NACE ou CITI car il existe une version NAF au format RDF avec le support des langues française et anglaise.

NAF a été créée en 1993 ; sa dernière version, NAF rév.2, date du 1er janvier 2008. Elle présente une structure arborescente à cinq niveaux : 21 sections, , 88 divisions 272 groupes, 615 classes, 732 sous classes. En analysant la nomenclature de NAF, nous avons constaté que la section J intitulée « Information et Communication » correspond au domaine de l'informatique. Dans cette section, nous trouvons 6 divisions : (i) Edition, (ii) Production de films cinématographiques, de vidéo et de programmes de télévision ; enregistrement sonore et édition musicale, (iii) Programmation et diffusion, (iv) Télécommunication, (v) Programmation, conseil et autres activités informatiques et (vi) Services d'information. Nous avons choisi de nous focaliser sur les trois dernières sections qui couvrent la majorité des données de Silex. Au final, notre ontologie NAF du domaine de l'informatique contient 30 concepts.

### 3.2.2 Kompass

Kompass [8] est la classification internationale d'activités la plus étendue du marché. Cette classification, dont la version d'origine a été créée en 1947, permet de classer les entreprises selon les produits et services qu'elles fournissent. Début 2014, une nouvelle version a été mise en exploitation, nommée WF13. Elle intègre des activités très récemment apparues sur le marché et propose une nouvelle structure hiérarchique qui prend en compte les dernières évolutions des différents secteurs d'activités de l'économie mondiale. WF13 propose un classement de 55 000 produits et services présentés en une arborescence de 4 niveaux : (i) 15 familles, (ii) 67 secteurs (iii) 3014 branches et (iv) 55450 produits et services [9]. L'inconvénient

6. http ://cigref.hr-ingenium.com/accueil.aspx

7. https ://www.insee.fr/fr/information/2406147

8. http ://www.kompass-international.com/Corporate/home.html

9. http ://www.kompass-international.com/Corporate/home/kompass-know-how/processing-the-data/classification.html

de cette classification est qu'elle est exposée via un site web uniquement. Nous avons extrait cette classification en utilisant un crawler pour la transformer au format SKOS-RDF. En nous focalisant sur le domaine informatique, nous avons identifié la famille « Informatique, Internet et R&D » qui contient 3 secteurs : (i) Informatique et Internet, (ii) Architectes, bureaux techniques et sociétés de conseil en ingénierie et (iii) Recherche et essais. L'ontologie Kompass contient 1370 concepts.

## 4 Construction des ontologies internes à Silex

En comparant les référentiels internes de Silex et les ressources définis précédemment, nous avons pu identifier des nouveaux concepts spécifiques à Silex. Afin de profiter de la richesse de ces référentiels internes, nous avons décidé de construire des ontologies internes.

### 4.1 Construction d'une ontologie des compétences interne à Silex

Le référentiel des compétences de la plateforme Silex est stocké dans une base de données et est enrichi par les utilisateurs. Nous sommes partis de ce fichier texte contenant 8470 termes afin de construire un thesaurus au format SKOS-RDF. Nous avons fait face à deux difficultés : (i) le référentiel mélange des termes appartenant à différents champs sémantiques comme les compétences, les métiers, les domaines d'activités, les villes, les pays, les langues, et (ii) le référentiel contient des termes composés, en langues française et anglaise, avec des fautes d'orthographe et des abréviations. Nous avons donc commencé par une étape de normalisation afin de supprimer tous les doublons et de regrouper les synonymes. Nous avons ainsi obtenu 6479 termes.

Nous avons ensuite utilisé la méthode de regroupement (*clustering*) hiérarchique en utilisant le word embedding (Mikolov *et al.*, 2013) et la métrique de similarité cosinus (Singhal *et al.*, 2001) pour identifier les groupes de termes relativement homogènes. Afin d'évaluer la qualité du résultat de ce regroupement, nous avons considéré les ontologies définies dans la section précédente et le résultat du regroupement comme des arbres phylogénétiques pour pouvoir ensuite les comparer. Nous avons considéré et implémenté trois métriques d'évaluation issues de la littérature : (i) la distance de Robinson-Foulds (Robinson & Foulds, 1981) permet de calculer la dissimilitude entre les arbres phylogénétiques en comptant les partitions qui n'existent que dans l'une des deux arbres, (ii) Cousin Pairs Distance (Panzetta, 2016) permet de détecter les phénomènes fréquents dans un arbre en cherchant le lien de parenté entre les nœuds, et (iii) Maximum Agreement SubTree (MAST) (Amir & Keselman, 1997) permet de définir l'arbre d'accord maximum comportant le plus grand nombre de branches. Cette étape d'évaluation est en cours.

### 4.2 Construction d'une ontologie des domaines d'activités interne à Silex

Le référentiel du domaine d'activités de Silex est stocké dans la base de données et élaboré par les commerciaux. Nous avons extrait ce référentiel au format CSV et construit un thésaurus au format SKOS-RDF contenant 14 concepts.

## 5 Alignement d'ontologies

Après la construction des ontologies, nous avons travaillé sur l'alignement entre les différentes taxonomies et ontologies utilisées. L'objectif est de tisser des liens entre : (i) les différents concepts de l'ontologie des compétences et de métiers, (ii) les ontologies du domaine d'activité et finalement (iii) entre les deux types d'ontologie. L'intérêt de construire ces alignements est de trouver les liens sémantiques entre les différents concepts (compétence, métier, domaine d'activité). L'alignement a été fait manuellement et l'évaluation de l'alignement a été faite par des experts chez Silex. La figure 1 explique l'approche d'alignement utilisée. Nous avons utilisé les propriétés `skos:broadMatch` pour définir une relation de généralisation entre deux concepts, `skos:exactMatch` pour définir un niveau de similarité élevé

entre deux concepts (mêmes labels), `skos:closeMatch` pour exprimer que deux concepts sont suffisamment similaires (labels différents), et la propriété `dcterms:references` [10] pour exprimer une référence à une ressource connexe.

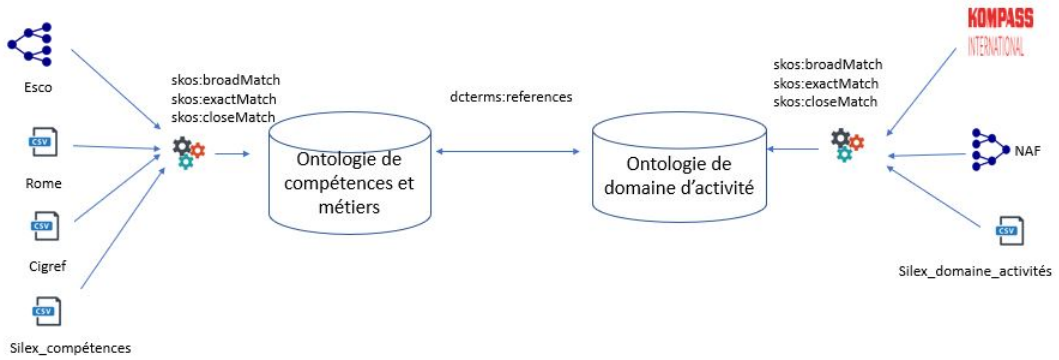



FIGURE *1 – Approche d'alignement des ontologies*

### 5.1 Alignement d'ontologies des compétences et des professions

Pour aligner les référentiels des métiers retenus, nous avons considéré ESCO comme référence en raison de son caractère multilingue et de sa complétude par rapport aux autres. Trois phases d'alignement ont été conduites : (i) entre ESCO et ROME, (ii) entre ESCO et Cigref, et (iii) entre ESCO et Silex_compétences. Le processus de l'alignement repose principalement sur la comparaison des labels préférés et des labels alternatifs des concepts.

Comme nous l'avons indiqué dans la présentation de ces ontologies, il existe différents niveaux de structuration de ces ontologies. Les niveaux les plus hauts dans les ontologies peuvent être vus comme des domaines, alors que le niveau le plus bas représente plutôt les métiers. Nous avons commencé par l'alignement des métiers (bas niveau) en cherchant la correspondance entre les noms et, par la suite, nous avons aligné les domaines (haut niveau). Pour aligner les domaines des ontologies, nous avons vérifié les concepts qui représentent les métiers et pour lesquels nous avons déjà établi un lien entre l'ontologie source et l'ontologie cible. Nous avons défini la règle suivante : s'il existe une correspondance entre le métier de l'ontologie source et le métier de l'ontologie cible, nous établissons une correspondance entre le domaine de l'ontologie cible et le domaine de l'ontologie source en utilisant la propriété `dcterms:references`. Nous avons ainsi pu aligner 67 concepts ESCO/ROME et 36 concepts ESCO/Cigref. L'alignement entre ESCO et Silex_compétences est en cours.

### 5.2 Alignement d'ontologie des secteurs d'activité

Nous avons considéré NAF comme l'ontologie de référence du secteur d'activité, et l'alignement consiste donc à retrouver une correspondance entre ses concepts et les concepts de Kompass en nous basant sur leur labels. Nous avons choisi de considérer niveau le plus bas des deux hiérarchies pour avoir la meilleure précision possible. Par exemple, la division 62 et le groupe 62.0 de NAF ont le même nom « Programmation, conseil et autres activités informatique ». Nous avons donc aligné le groupe 62.0 de NAF avec la branche de Kompass 57830 « Audit et conseil informatiques ». Nous avons aussi utilisé la propriété

10. https ://terms.tdwg.org/wiki/dcterms :references

`skos:narrowMatch` pour établir une relation de généralisation entre le concept « Programmation informatique » de NAF et les concepts « Logiciel de langage et de programmation » et « Services de programmation informatique » de Kompass. Nous avons ainsi défini 11 correspondances.

### 5.3 Alignement entre l'ontologie des compétences et des professions et l'ontologie des secteurs d'activité

La dernière étape du processus d'alignement est l'élaboration des liens entre l'ontologie des compétences et professions, et l'ontologie des secteurs d'activité. Cela correspond à aligner les deux ontologies de base qui sont respectivement ESCO et NAF. Pour cela, nous nous sommes basés sur un document fourni par Pôle Emploi qui permet de faire la correspondance entre ROME et NAF (pôle emploi, 2017). Ayant déjà établi les liens entre ESCO et ROME, nous déduisons ainsi les liens entre ESCO et NAF par transitivité. Par exemple, nous avons déjà aligné le concept ESCO ayant comme label « Programmeurs d'applications » avec le concept de ROME dont le label est « Études et développement informatique ». Le document de pôle emploi établit une correspondance entre ce dernier concept et les divisions 62 et 63 de NAF.De ce fait, nous avons défini une correspondance entre le concept d'ESCO et les divisions 62 et 63 de NAF. La figure 2 présente un exemple d'alignement entre ESCO et NAF basé sur ROME.

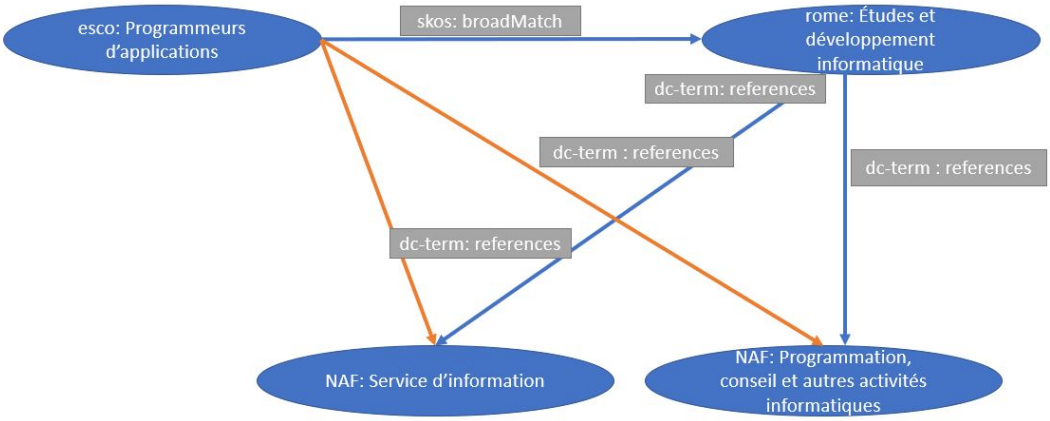



FIGURE 2 – *Exemple d'alignement ESCO-NAF basé sur ROME*

## 6 Conclusion et perspectives

Dans cet article, nous avons décrit notre approche de la construction d'une ontologie pour supporter une nouvelle approche du *sourcing*. Cette ontologie doit permettre d'annoter sémantiquement les descriptions textuelles des entreprises et leurs offres et demandes de prestations selon trois types de connaissances : les compétences, les professions et les domaines d'activités, dans le but d'améliorer la recommandation automatique des prestataires. Notre modélisation est basée d'une part sur une approche descendante de réutilisation des référentiels comme ESCO, ROME, NAF et Kompass, et d'autre part sur une approche ascendante de construction d'une ontologie à partir des données internes de l'entreprise. La dernière étape de la construction de notre ontologie consiste à aligner les référentiels, pour l'instant manuellement, afin de produire une ontologie la plus riche possible. A ce stade, notre ontologie ne couvre que le domaine de l'informatique. Le tableau 1 récapitule les différentes ontologies réutilisées ou construites pour représenter les données de Silex pour le domaine de l'informatique. Comme perspective immédiate, nous envisageons d'étendre notre ontologie à d'autres

| Ontologie | Type d'ontologie | Nombre de concepts | Langues | Format | Nombre d'alignement |
|---|---|---|---|---|---|
| ESCO | compétences et professions | 167 professions ; 524 compétences | 26 langues | SKOS-RDF | 34 liens avec NAF |
| ROME | compétences et professions | 118 | français | transformation en SKOS-RDF | 67 liens avec ESCO |
| Cigref | compétences et professions | 42 | français | transformation en SKOS-RDF | 36 liens avec ESCO |
| Silex_compétence | compétences et professions | | français anglais | transformation en SKOS-RDF | en cours |
| NAF | domaines d'activités | 30 | français anglais | SKOS-RDF | 34 liens avec ESCO |
| Kompass | domaines d'activités | 1370 | français | transformation en SKOS-RDF | 11 liens avec NAF |
| Silex_activité | domaines d'activités | 14 | français | transformation SKOS-RDF | 7 liens avec NAF |

TABLE *1 – Récapitulation des différentes ontologies réutilisées ou construites pour représenter les connaissances de Silex pour le domaine de l'informatique*

domaines tels que le marketing et les services généraux, en visant autant que possible une automatisation de la phase d'alignement.

Par ailleurs, nous avons amorcé un travail de catégorisation automatique des textes décrivant les entreprises, offres et demandes de services, dans le but à terme d'apparier automatiquement offres et demandes.

## Références


AMIR A. & KESELMAN D. (1997). Maximum agreement subtree in a set of evolutionary trees : Metrics and efficient algorithms. *SIAM Journal on Computing*, **26**(6), 1656—-1669.

AMOURACHE F., BOUFAÏDA Z. & YAHIAOUI L. (2008). Construction d'une ontologie basée compétence pour l'annotation des cvs/offres d'emploi. In *10th Conference on Software Engineering and Artificial Intelligence (MCSEAI), Maghrebian Conference on Information Technologies (28-30 april)*, p. 1–7.

ESCHENLAUER R. (2013). Le sourcing.

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

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv :1301.3781*.

NOY N. F. & MCGUINNESS D. L. (2000). Développement d'une ontologie 101 : Guide pour la création de votre première ontologie. *Université de Stanford, Stanford, CA, 94305. Traduit de l'anglais par Anila Angjeli, BnF, Bureau de normalisation document*.

NOY N. F., MCGUINNESS D. L. *et al.* (2001). Ontology development 101 : A guide to creating your first ontology.

PANZETTA A. (2016). A preliminary study on a new similarity measure for phylogenetic trees. Master's thesis, Università Ca' Foscari Venezia.

PÔLE EMPLOI (2017). Correspondance naf 2008-rome v3.

ROBINSON D. R. & FOULDS L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, **53**, 131–147.

SINGHAL A. *et al.* (2001). Modern information retrieval : A brief overview. *IEEE Data Eng. Bull.*, **24**(4), 35–43.

SUÁREZ-FIGUEROA M. C. & GÓMEZ-PÉREZ A. (2009). Neon methodology for building ontology networks : a scenario-based methodology. In *Proceedings of the International Conference on Software, Services & Semantic Technologies* : Sofia.

TCHOUANTO POOSIA P. (2014). Modularisation des ontologies. Master's thesis, Université du Québec à Montréal.

USCHOLD M. & GRUNINGER M. (1996). Ontologies : Principles, methods and applications. *The knowledge engineering review*, **11**(2), 93–136.

USCHOLD M. & KING M. (1995). Towards a methodology for building ontologies. In *Workshop on Basic Ontological Issues in Knowledge Sharing, held in conjunction with IJCAI-95*.

# Construction d'ontologies : des fondements théoriques à la mise en œuvre

La première contribution de cette session nous renvoie aux origines métaphysiques du terme ontologie. Elle propose un cadre conceptuel pour représenter différentes catégories ontologiques, en particulier les entités "occurrentes" et les "continuantes". La deuxième contribution ambitionne d'outiller les différentes étapes de construction des ontologies et pour cela propose une description ontologique de ces outils afin de faciliter leur sélection en fonction de différents critères. Enfin, nous nous intéresserons aux propriétés qu'un langage visuel de représentation des connaissances doit satisfaire. Ainsi ces contributions ambitionnent, chacune à sa façon, de caractériser certains traits d'une ontologie pour nous permettre de mieux les appréhender.

# Une alternative à la distinction 'continuant' vs 'occurrent'

Gilles Kassel

Laboratoire MIS, Université de Picardie Jules Verne
33 rue Saint-Leu, 80039 Amiens Cedex 1
Gilles.kassel@u-picardie.fr

**Résumé** : Nous posons les bases d'une ontologie de haut niveau de particuliers dont les principes de structuration sont radicalement différents de l'opposition 'continuant' vs 'occurrent' classiquement retenue en Ontologie Appliquée. Ces principes découlent d'une analyse nouvelle de l'ontologie des entités « survenantes » ou « occurrentes ». L'analyse intègre d'une part des travaux récents en ontologie des processus, rapprochant ces derniers des objets, en les assimilant à des substances continuantes (Stout, Galton). Elle propose par ailleurs de distinguer nettement *processus* et *événements*, identifiant ces derniers à des objets abstraits de pensée côtoyant les propositions. Enfin, nous ouvrons notre inventaire ontologique aux *propriétés* et aux *faits* dont l'existence réelle est admise (Fine, Armstrong). Ce cadre permet de rendre compte tout à la fois du monde physique dans ses aspects statiques et dynamiques et de la façon dont des agents conçoivent son histoire, en faisant jouer aux primitives le rôle suivant : les *faits* rendent compte de la vie des substances – *objets* et *processus physiques* – tandis que les *événements* rendent compte, pour des sujets cognitifs, de l'histoire de la vie de ces substances.

**Mots-clés** : Objet, Processus, Événement, Propriété, Fait, Proposition, Continuant, Occurrent

## 1 Introduction

Les catégories de *processus* et d'*événement* paraissent avoir un destin intimement lié dans les théories métaphysiques courantes. Toutes deux incarnent la dynamicité du monde en correspondant à des entités qui « occurrent » ou « surviennent ». Elles s'opposent aux *objets* et à la *matière* dont ceux-ci sont constitués, ces entités incarnant la stabilité du monde en « endurant » ou « continuant ». Une doctrine solidement ancrée considère la relation entre processus et événement comme analogue à celle entre matière et objet : les processus sont la « matière » des événements, ce qui revient à dire que les événements sont « constitués » de processus (Mourelatos, 1978)(Galton & Mizoguchi, 2009)(Crowther, 2011)(Steward, 2013). Selon cette conception, processus et événements sont des entités concrètes habitant une même région spatio-temporelle du monde physique. Cette doctrine ne semble toutefois pas être gravée dans le marbre. De nouvelles propositions dans plusieurs domaines de la métaphysique la remettent même en cause.

Quelques travaux récents sur la métaphysique des processus, notamment, assimilent ces derniers plutôt à des entités portant temporairement des propriétés et capables de changer à la manière d'endurants (Stout, 1997, 2003)(Galton, 2006). Bref, il s'agirait de continuants occurrents « survenant » (Stout, 2016), à moins qu'il ne s'agisse d'occurrents continuants, c'est-à-dire d'entités qui, tout en étant étendues dans le temps et ayant des parties temporelles, auraient temporairement des propriétés et seraient susceptibles de changer (Steward, 2013,

2015). Un enjeu, lié à la caractérisation de la nature des processus, est ainsi la question de l'ontologie du temps et de l'occupation du temps (Crowther, 2011). Un enjeu corrélé est de préciser la notion de « survenue » par rapport à celle d'« existence » lorsque ces notions se rapportent à des occurrents.

Parallèlement, du côté de la métaphysique des événements, même si la conception de (Davidson, 1969) d'événements en tant que des particuliers concrets continue de tenir le haut du pavé, la littérature n'en finit pas d'exprimer des interrogations. Certains auteurs ont très tôt exprimé leur scepticisme quant à l'existence d'événements (Horgan, 1978)(Hacker, 1982a), en tout cas d'événements tels que définis par Davidson. Ces doutes ont été relayés récemment par des auteurs travaillant sur l'ontologie de l'action (les actions étant réputées être une espèce d'événements), du fait de difficultés à articuler la conception Davidsonnienne des événements avec notre connaissance de la phénoménologie des actions (Hornsby, 2012)(Steward, 2012). Par ailleurs, des auteurs admettant la co-existence de processus et d'événements sont tentés d'identifier ces derniers à des entités abstraites (Gill, 1993), reprenant une proposition historiquement faite par Chisholm (1970) ou Wilson (1974). De fait, si l'existence de deux catégories ontologiques distinctes - *processus* et *événements* – ne semble pas remise en cause et s'avère même utile (Steward, 2015), en tout cas des doutes existent sur le bienfondé de la doctrine de la constitution d'événements par des processus.

Notre objectif dans cet article est justement de proposer un cadre ontologique qui soit suffisamment cohérent pour qu'il représente une alternative crédible à cette thèse de la constitution. Notre proposition repose tout autant sur une démarche que sur un cadre conceptuel organisant différemment les oppositions entre catégories ontologiques. Elle doit son origine à une proposition d'Antony Galton (2006, 2008) de substituer à la distinction 'continuant' vs 'occurrent' la distinction EXP vs HIST (Fig. 1) entre le monde tel qu'il se déroule et son histoire (Galton 2008, p. 323) :

> *[…] processes differ markedly from events in their relation to change. Whereas events are fixed items of history which cannot be described as undergoing change, processes are more like ordinary objects in that they can be directly present at one time and can undergo change as time proceeds. This leads to a fundamental ontological distinction between EXP, the dynamic experiential world of objects and processes as they exist at one time, and HIST, the static historical overview populated by events that are generated by the ongoing process in EXP.*

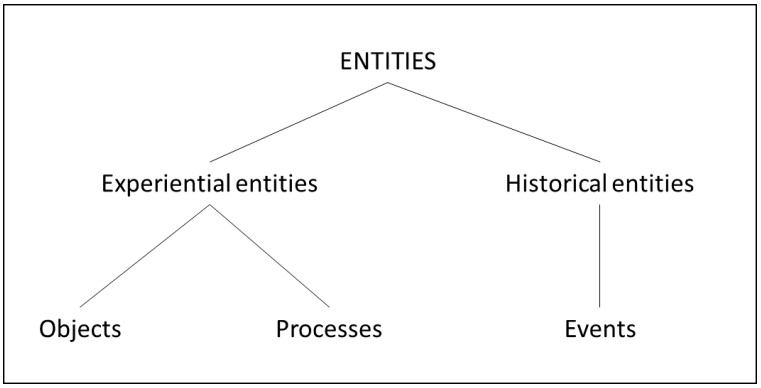



FIGURE 1 – La distinction entre entités *expérentielles* et *historiques* (tiré de (Galton, 2008))

Dans cet article, nous reprenons à notre compte cette distinction entre EXP et HIST, toutefois, là où Galton se contente d'y voir une différence de perspectives de description du monde, nous proposons de l'identifier à une distinction métaphysique radicale. La thèse que

nous soutenons est que les objets et processus physiques existent dans le monde et ont une vie indépendamment de notre façon de les penser tandis que les événements sont des objets abstraits de pensée, des construits, dont la fonction cognitive essentielle est de disposer d'histoires du monde. Pour résumer notre point de vue, en redéfinissant les catégories introduites par Galton : EXP est le monde physique en mouvement tandis que HIST en est son histoire (passée, présente et à venir) construite par des sujets humains.

Cette thèse est tout à la fois le fruit d'une démarche de recherche métaphysique spécifique et une réponse à un questionnement exprimé par plusieurs auteurs (dont Galton lui-même) quant aux modes d'existence contrastés entre processus et événements, quand ce questionnement ne remet pas en cause l'existence même des événements.

Notre démarche suit le travail méthodologique d'enquête auquel nous convie Maurizio Ferraris (2014) pour traquer la frontière entre faits et interprétations (pp. 72-74) :

> *Le point n'est donc pas d'affirmer qu'il y a une discontinuité entre les faits et les interprétations, mais plutôt de comprendre quels sont les objets construits et quels sont ceux qui ne le sont pas […] Ce travail consiste à faire une distinction méticuleuse entre l'existence des choses qui n'existent que pour nous – des choses qui n'existent que s'il y a une humanité – et des choses qui existent même en l'absence de l'humanité.*

Ce travail d'enquête nous conduit à positionner les processus physiques du côté des choses existant même en l'absence de l'humanité et les événements du côté des construits. Plus spécifiquement, le cadre ontologique auquel nous parvenons est le suivant :

- Le monde physique est peuplé de substances – *objets* et *processus* – qui, en endurant, assurent sa stabilité tout autant que sa dynamicité ;
- Ces substances, portant temporairement des propriétés et entretenant des relations avec d'autres substances, ont une vie, laquelle consiste en des *faits* existant dans le monde physique ;
- Des sujets cognitifs, plongés dans le monde physique, se représentent au moyen d'*événements* l'histoire passée, présente et future du monde pour interagir avec lui.

Dans l'article, nous commençons par présenter en Section 2 notre cadre ontologique. La présentation reprend des éléments publiés dans (Kassel, 2017) en élargissant le cadre à la catégorie des *faits*. Ceci permet de distinguer clairement, d'une part, la *vie* des substances et, d'autre part, l'*histoire* de cette vie. Enfin, en Section 3, nous discutons l'opposition entre 'continuants' et 'occurrents' et montrons qu'elle devient caduque pour organiser le haut niveau d'une ontologie de particuliers.

## 2 Notre cadre ontologique

Avant de mettre en scène nos primitives ontologiques, nous précisons le type d'ontologie que nous cherchons à établir d'où dérivent des contraintes quant au choix des primitives.

### 2.1 Des principes de choix de nos primitives ontologiques

Comme point de départ, indiquons que nous suivons le projet de Peter Strawson (1959) d'établir une métaphysique *descriptive* visant à décrire « *the actual structure of our thought about the world* ». L'objectif est ainsi d'établir des catégories et notions rendant compte de la façon dont nous concevons le monde et, adoptant une perspective contemporaine de l'ontologie, nous retenons deux modes principaux de structuration du monde. D'une part, nous retenons que

notre appréhension du monde nous conduit à distinguer trois types de réalité : physique, mentale et sociale. Ce découpage de la réalité, adopté par les « nouveaux réalistes » comme Ferraris, relève de la reconnaissance de *modes d'existence* distincts. Un second découpage s'appuie sur la notion de *niveau d'abstraction* : si nos conceptions de sens commun nous révèlent des entités à un niveau mésoscopique, des conceptions savantes instruites par des sciences nous révèlent des entités situées à un niveau plus élevé (ex : astronomique) ou plus bas (ex : microscopique). Les liens entre ces niveaux tiennent davantage de la relation de *constitution* que de la relation *tout/parties* (Masolo, 2010). L'important, pour la caractérisation de nos catégories et notions, est de respecter ces niveaux d'abstraction. Ainsi, si une table ou tout autre objet matériel nous apparaît au niveau mésoscopique comme un objet, le contenu de la même région spatio-temporelle peut être décrit au niveau nano comme un tourbillon de particules dans du vide. La description au niveau nano révèle d'autres entités et ce n'est pas pour autant qu'une table ou tout autre objet matériel doit être identifié à une « masse dynamique » (nous nous démarquons ici de métaphysiciens soutenant des ontologies mono-catégorielles priorisant la primitive du processus, comme on peut le voir chez Seibt (2008)).

Pour le choix de nos primitives ontologiques, une contrainte que nous posons est leur utilité à rendre compte du plus grand nombre possible de niveaux. Ce critère qualifie en priorité l'*objet physique* dont la présence est attestée à pratiquement tous les niveaux (excepté sans doute au niveau nano) aussi bien dans les théories cognitives de sens commun qu'en physique et dans la théorie de la mesure. Suivant la conception que nous allons donner du *processus physique*, en proposant dans la section suivante une relation duelle entre processus et objets physiques, cette primitive bénéficiera de la même robustesse.

### 2.2 Les substances

Pour débuter notre inventaire, nous adoptons une conception classique de l'objet physique. Un *objet physique* matériel est quelque chose :

o_i) existant à des temps ;
o_ii) ayant des propriétés à des temps (ex : couleur, odeur, masse, volume) ;
o_iii) pouvant changer dans le temps.

Ces propriétés traduisent la conception d'une entité 3D existant *à* des temps et endurant *dans* le temps en ayant des propriétés pouvant varier dans le temps. Les propriétés (o_ii) et (o_iii) caractérisent la vie de l'objet physique : nous y revenons en §2.3. En extension, des exemples sont des objets maximalement connectés, qu'ils soient inertes (ex : une pierre, une pomme détachée d'un arbre, une molécule d'eau, une planète), animés-vivants (ex : un être humain, une fleur, un arbre) ou artefactuels (ex : une chaise, un presse-papier, une télévision). Dans le cas des artefacts, nous verrons en §2.3 qu'il s'agit d'objets physiques « simpliciter » dotés d'une vie sociale.

La caractérisation que nous donnons maintenant des processus physiques est largement fondée sur l'idée de continuants dynamiques développée principalement par Rowland Stout (1997, 2003) et Antony Galton (2006, 2008). Un *processus physique* est quelque chose :

p_i) existant à des temps ;
p_ii) ayant des propriétés à des temps (ex : direction, vitesse d'exécution, niveau sonore, amplitude spatiale) ;
p_iii) pouvant changer dans le temps.

Nous retrouvons la même caractérisation que pour l'objet physique, un ensemble de trois propriétés définissant ce que nous considérons être une *substance* dans cet article. En extension, des exemples de processus physiques sont : le mouvement d'un objet physique (conduisant au déplacement ou à la rotation de l'objet sur lui-même) ; la croissance en taille d'un corps physique ; le processus de vie d'une personne ; le mûrissement d'un fruit ; l'oxydation d'un objet métallique ferreux ; la fonte d'un glacier.

Reprenons les propriétés attribuées aux processus physiques une à une pour les éclairer. On trouve la première (p_i) exprimée chez Stout (1997, p. 26) :

> *The phrase, 'What is happening now', is naturally taken to denote a whole process ; and we do want to claim that what is happening now is literally identical with what is happening at some other time – the very same process.*

Cette thèse repose sur un premier engagement fort quant à la nature des processus physiques : la cause fait partie du processus, lequel processus constitue le moteur de déplacements et, plus généralement, de changements. En ce sens, on peut dire que le déplacement d'un objet – un changement de position – résulte d'un mouvement de l'objet, à savoir le processus. Le fait d'associer la cause au processus, toujours selon Stout, est nécessaire pour tordre le cou à la conception Russellienne d'un mouvement comme série d'états successifs (Stout, 2003)[1] :

> *[The] motion should not be understood in Russell's way as the arrow being in one state and then in another and in the meantime being in all the intervening states. The arrow's motion is what gets it through this continuous series of states - it effects the transition.*

La conception Russellienne est en fait héritière de la philosophie de l'école d'Élée qui, en concevant précisément des déplacements comme une série d'états, donc d'immobilités, a tourné le dos à une compréhension du mouvement comme un fluant « indivisible », pour reprendre la terminologie de Bergson (1934)[2].

Venons-en à la possibilité pour un processus de pouvoir porter des propriétés pouvant varier dans le temps, les propriétés (p_ii) et (p_iii). Il s'agit, selon Galton (2006, p. 6), d'une caractéristique rapprochant les processus des objets :

> *Like objects, processes can change: the walking can get faster, or change direction, or become limping. All around us processes undergo changes: the rattling in the car becomes louder, or change rhythm, or may stop, only to start again later. The flow of the river becomes turbulent; the wind veers to the north-west.*

On notera que le fait d'avoir identifié le processus au moteur du changement et non au changement lui-même est déterminant dans la caractérisation de continuant du processus. Les adeptes d'une conception du processus comme un changement *occupant* du temps avancent qu'un changement ne peut lui-même changer et que, dès lors, la seule substance existante est celle portant le processus (l'argument a été réitéré récemment par exemple par Thomas Crowther (2018)). Ainsi, si la vitesse du mouvement d'un corps vient à changer entre deux

[1] L'idée d'encapsuler la cause dans le processus n'est en fait pas nouvelle. On la retrouve en effet dans la notion de *comportement* de Fred Drestke (1988). Selon Drestke, un « comportement » ou « processus causal » consiste dans le fait qu'une cause provoque un mouvement. Le processus physique que nous caractérisons ici correspond ainsi au *comportement* ou *processus causal* de Drestke.

[2] Bergson attribue à cette école le fait que le processus ait été relégué au rang d'accident (1934, V) : « La métaphysique est née, en effet, des arguments de Zénon d'Élée relatifs au changement et au mouvement. C'est Zénon qui, en attirant l'attention sur l'absurdité de ce qu'il appelait mouvement et changement, amena les philosophes – Platon tout le premier – à chercher la réalité cohérente et vraie dans ce qui ne change pas ».

instants $t_1$ et $t_2$, ce changement est à attribuer non au mouvement mais à ce corps se mouvant. Par exemple, Paul possède la propriété de 'Marcher à la vitesse de 5 km/h' à $t_1$ et la propriété de 'Marcher à la vitesse de 6 km/h' à $t_2$. Au contraire, la conception que nous adoptons nous conduit à considérer qu'un processus de marche de Paul a une 'vitesse de 5 km/h' à $t_1$ et une 'vitesse de 6 km/h' à $t_2$.

Pour compléter la caractérisation de nos processus physiques, évoquons maintenant un engagement ontologique supplémentaire, à savoir le fait qu'un processus ne soit pas un continuant flottant dans l'air, mais soit ancré dans un *objet support* : il s'agit du mouvement d'une *flèche*, du mûrissement d'un *fruit*, de la fonte d'un *glacier*, etc. Pour rendre compte de ce lien fort constitutif, nous reprenons à notre compte la relation d'*énaction* introduite par Galton et Mizoguchi (2009). Pour ces auteurs, dire qu'un objet « énacte » un processus revient à dire qu'un objet porte un processus « externe » ou « comportement » (p. 94) :

> *The key notion is that an object, considered from a particular point of view, is characterized in terms of the processes it enacts. These are what we call the external processes or behavior of the object. This behavior arises as a result of various internal processes which causally contribute to it.*

Cette caractérisation de la relation d'énaction par Galton et Mizoguchi repose sur une conception de l'objet comme interface entre des processus internes et externes. Elle revient à introduire une relation duelle de dépendances qu'expriment les propriétés suivantes :

- o_iv) l'intégrité (existence) de tout objet physique est maintenue par des processus physiques ;
- p_iv) tout processus physique est énacté par un objet physique.

De ces propriétés, nous pouvons déduire que l'existence d'un objet physique dépend de celle d'autres objets situés à des niveaux (ontologique ou d'abstraction) inférieurs et, de façon duale, que l'existence d'un processus physique dépend de celle d'autres processus situés à des niveaux inférieurs. Ainsi, un processus de marche d'une personne n'est possible que si, notamment et à un premier niveau, des processus énactés par ses jambes existent concomitamment, et que si, notamment et à un niveau inférieur, des processus physiologiques énactés par ses organes existent concomitamment. Signe de cette relation duelle, l'objet et le processus physique sont habituellement crédités d'une vie continue. Par exemple, rien dans la notion de processus de marche n'évoque une fin possible : celle-ci peut se produire contingentement par exemple suite à une fatigue de la personne ou suite à une décision de la personne.

### 2.3 La vie des substances

Notre inventaire ontologique est à ce stade composé de substances, des objets et des processus physiques. S'agissant d'entités persistant dans le temps tout en conservant leur identité, elles existent à différents instants et ceci nous conduit à parler de leur vie. La vie de ces entités est limitée dans le temps entre le moment où elles acquièrent une existence et le moment où elles cessent d'exister. Certaines ont une vie brève, à l'échelle humaine, comme ces particules élémentaires étudiées par les physiciens des hautes énergies dont la durée de vie (et celle de leurs mouvements) n'est qu'une infime fraction de seconde. D'autres ont une vie longue voire semblant éternelle, toujours à l'échelle humaine, à l'instar des astres et galaxies et de leurs mouvements.

Dans cette section, nous donnons corps au concept de *vie* des objets et des processus physiques en lui conférant une certaine légitimité ontologique. Sans pour autant en faire une catégorie ontologique à part entière, nous précisons cette notion en la rattachant à deux

primitives ontologiques ayant déjà acquis leurs lettres de noblesse en ontologie formelle (même si plusieurs théories existent à leur sujet), celle de *propriété* et celle de *fait*.

Pour débuter par les propriétés, une clarification terminologique s'impose : le terme « propriété » est communément utilisé pour dénoter, d'une part, des universaux (types) et des tropes (instances) attachés à des substances (Armstrong, 1997) et, d'autre part, des concepts ou catégories qui structurent nos représentations et théories du monde (Margolis & Laurence, 1999). Dans cet article, nous réserverons le terme « propriété » pour désigner la première catégorie d'entités. D'importance pour notre propos, on notera que certaines propriétés sont « naturelles » (nous les qualifierons de « physiques ») en étant indépendantes de toute pensée humaine (ex : 'Être un galet', 'Être une molécule d'eau') tandis que d'autres sont « sociales » car n'existant que parce qu'elles sont pensées par un agent ou une communauté d'agents (ex : 'Être un presse-papier', 'Être un billet de banque')(Searle, 1995). Cette distinction se reflète dans la façon dont les objets « portent » ou « exemplifient » des propriétés. Elle a également à voir avec une distinction concernant les objets eux-mêmes entre objets « physiques » et « non physiques ».

Le lecteur trouvera dans (Armstrong, 1997, ch. 3 et ch. 4) un large traitement des propriétés physiques dont l'existence est supposée. Ces propriétés correspondent à des « manières d'être » d'entités substantielles. Différentes théories ont été proposées pour rendre compte de leur nature. Pour certains théoriciens, les propriétés sont des universaux, autrement dit des entités se répétant à l'identique dans des substances ; d'autres théoriciens considèrent que les propriétés sont des tropes, autrement dit des particuliers inhérents à leur porteur. Dans cet article, nous ne prendrons pas parti. Précisons que des conceptions hybrides combinant universaux et tropes ont été proposées dans la littérature. En Ontologie Appliquée, les propriétés faisant l'objet d'une expérience perceptuelle de la part de sujets (ex : couleur, forme, odeur, masse) sont souvent conçues comme des tropes. On trouve un tel traitement dans BFO (Grenon & Smith, 2004) et DOLCE (Masolo *et al.*, 2003).

Concernant les propriétés sociales, celles-ci participent de la construction d'une partie du monde appelée « réalité sociale » (Searle, 1995, 2010). Pour notre propos, nous nous contenterons de considérer, en empruntant la terminologie de Amie Thomasson (2003), que certaines propriétés sociales portent sur des objets physiques (typiquement, des propriétés fonctionnelles comme 'Être un presse-papier') en en faisant des objets « sociaux concrets », tandis que d'autres portent sur des objets non physiques (ex : 'Être une monnaie', 'Être une loi', 'Être une nation'), en en faisant des objets « sociaux abstraits ». Ces derniers correspondent à la catégorie des 'Non-Physical objects' introduite dans DOLCE pour rendre compte des objets n'existant que parce que des agents les conçoivent et communiquent à leur propos. On notera que cette distinction *physique* vs *non physique* tient également pour les processus. Parmi les processus non physiques nous pouvons distinguer des changements de propriétés sociales pour des objets physiques (ex : 'devenir veuf pour une personne', 'perdre de sa valeur faciale pour un billet de banque') ou pour des objets non physiques (ex : 'se déprécier pour une monnaie', 'tomber en désuétude pour une loi'). Nous en profitons pour noter que, dans cet article, nous ne traiterons pas de processus non-physiques.

La catégorie de propriété étant introduite, nous continuons à étendre notre inventaire en admettant cette fois celle de *fait* pour rendre compte d'entités comme 'Paul est embarrassé' ou 'Paul est à côté de Marie'. La thèse de l'existence de telles entités, avancée par de nombreux philosophes (ex : Kit Fine (1982), Donald Armstrong (1997)), est une thèse compagnon de la théorie réelle des propriétés : l'existence simultanée de la substance 'Paul' et de la propriété 'Être embarrassé' ne signifie pas pour autant que la substance 'Paul' exemplifie à un temps

donné la propriété ‘Être embarrassé’, le même constat tenant pour la substance ‘Paul’ et la relation ‘Être à côté de Marie’. Le fait, ou la « circonstance » (pour reprendre le terme de Fine), correspond à ce lien interne unissant substance et propriété/relation en une entité à part entière. L’argument principal de l’existence des faits est qu’ils constituent un vérifacteur (truth maker), autrement dit ce qui rend vrai dans le monde des propositions comme ‘Paul est embarrassé’ ou ‘Paul est à côté de Marie’ (Armstrong, 1997).

Le lecteur aura noté l’embarras notationnel dans lequel on se trouve quand on se pose la question *Que sont les faits ?* ou, comme Andrea Iacona (2013), *Que sont les propositions ?*. On a vite fait d’élaborer un discours du type « il est vrai que le ciel est bleu car le ciel est bleu », voulant signifier que « la *proposition* selon laquelle ‘le ciel est bleu’ est vraie car le *fait* ‘le ciel est bleu’ existe ». Pour résoudre cet embarras, dans cet article nous adoptons une notation commune (pour les philosophes étudiant les faits) faisant apparaître la structure du fait en ses constituants : <Ciel, Être bleu>. Une même syntaxe a été proposée également dans la littérature pour les propositions et les événements, revenant à expliciter leurs constituants. Par ailleurs, pour lever toute ambiguïté lorsque nous parlons concomitamment de faits et de propositions, nous explicitons la catégorie d’entité en postfixant la notation d’un indice : $<>_F$, $<>_P$ (ex : <Paul, Être embarrassé$>_F$ tient pour le *fait*, tandis que <Paul, Être embarrassé$>_P$ tient pour la *proposition*).

De même que nous avons distingué supra entre propriétés physiques et sociales, il est important de distinguer plusieurs types de faits. Tels que nous les avons introduits, les faits/circonstances concernent des substances (objets et processus physiques) et il est communément admis que le lien attachant la substance à une propriété correspond à une instanciation dénotée par « est » : Paul *est* ‘Être embarrassé’ ; Paul *est* ‘Être à côté de Marie’. Pour ces faits physiques, une conception courante est par ailleurs de considérer le temps comme constituant : <Paul, Être embarrassé, À l’instant$>_F$, <Paul, Être à côté de Marie, À l’instant$>_F$. La raison en est que les substances ont leurs propriétés *à* des temps (cf. les propriétés (o_ii) et (p_ii)). Par contre, toujours pour ce qui concerne les substances, il est d’autres faits pour lesquels l’association d’une propriété à une substance correspond à une stipulation humaine pouvant reposer sur une convention sociale (pour reprendre l’analyse de Searle (1995)) : ce morceau de papier *compte pour* ‘Être un billet de banque de 10 euros’ (pour un agent ou une communauté d’agents, et dans certaines circonstances) ; Ce galet *compte pour* ‘Être un presse-papier’[3]. Les substances communément considérées en exemples, aussi bien par Searle et par Thomasson, sont des objets physiques. Maintenant, le fait d’admettre les processus physiques comme substances nous procure d’autres exemples : tel mouvement des lèvres *compte pour* un sourire ; telle élévation du bras *compte pour* une indication de tourner à gauche. La différence avec les faits précédents est que ces derniers reposent sur l’attribution de propriétés par des agents à des substances. Ce ne sont plus des « faits bruts » (pour reprendre la terminologie de Searle) mais des « faits sociaux » construits. Les conditions d’existence des faits dès lors varient, suivant qu’elles correspondent à une instanciation indépendante de toute pensée humaine, pour les faits bruts, ou à une stipulation humaine, pour les faits sociaux.

[3] Précisons que plusieurs théories ontologiques ont été proposées dans la littérature pour rendre compte de telles stipulations consistant à attribuer une fonction à un objet. Suivant la théorie des artefacts développée par Borgo *et al.* (2014), attribuer une fonction (ex : ‘Être un presse-papier’) à un objet physique (ex : un galet) revient à créer un second objet physique – l’artefact galet-presse-papier – constitué du premier objet physique. Au contraire, et en cohérence avec notre choix d’établir une ontologie descriptive, l’interprétation que nous donnons à l’acte cognitif est celle de l’ajout d’une propriété à ce même objet physique (Kassel, 2010).

Les exemples de faits que nous venons de prendre illustrent ce que nous entendons par « vie » d'une substance : il s'agit d'une collection de faits temporels se rapportant à la substance (pour lesquels la substance apparaît comme constituant). Les exemples que nous avons considérés concernent plutôt des objets. À ce propos, on notera que la propriété (p_iv) caractérisant les processus – le fait qu'un processus soit énacté par un objet à un temps – inscrit les processus dans la vie des objets (les objets ont une vie processuelle). Mais, intéressons-nous maintenant à la vie des processus. Une catégorie importante de faits participant de la vie des processus correspond à la *perpétuation* des processus. De tels phénomènes interviennent lorsque, par exemple, le mouvement d'une masse d'air « perpétue » le mouvement d'une feuille, le déplacement de la masse d'air « provoquant » le déplacement de la feuille ; ou bien, lorsque le mouvement d'un bras « perpétue » le mouvement d'une montre portée au poignet, une élévation du bras « provoquant » une élévation de la montre. Dans les descriptions que nous venons de donner, en évoquant, d'une part, des mouvements-processus « perpétuant » d'autres mouvements-processus et, d'autre part, des déplacements-événements « provoqués » par d'autres déplacements-événements, nous avons fait attention à distinguer les processus des événements. Les relations intitulées « perpétue » et « provoque » font partie d'un catalogue de relations causales entre états, processus et événements, tel que décrit par Galton (2012). Pour en rester aux processus (sachant que nous ne donnons notre définition des événements qu'en §2.4), nous retenons du catalogue de Galton la relation *perpetuatesAt(p,p',t)* signifiant que « le processus *p* perpétue (ou entretient) le processus *p'* au temps *t*). Cette relation tient entre deux processus déjà existants et on peut la comprendre comme une propagation de causalité : un processus énacté par un objet entretient (en participant de sa cause) un autre processus énacté, soit par un autre objet (cf. nos exemples), soit par le même objet (ex : le mouvement d'un corps entretient l'échauffement de ce corps). Il est intéressant de noter que les faits de perpétuation représentent une partie de la dynamique du monde. Cette remarque met en avant un argument supplémentaire en faveur de l'existence des faits bruts ou physiques (venant compléter l'argument du vérifacteur).

## 2.4 L'histoire de la vie des substances

Venons-en à évoquer un construit humain, à savoir la façon dont les humains conceptualisent l'histoire de la vie des substances – processus et objets physiques – peuplant le monde. La principale catégorie ontologique que nous allons ajouter à notre inventaire est celle d'*événement*. Pour commencer à fixer les idées, précisons d'emblée que la notion d'événement que nous visons est proche de celles théorisées historiquement par Roderick Chisholm (1970) ou Neil Wilson (1974), faisant des événements des objets abstraits à côté des propositions. Par ailleurs, pour continuer à fixer les idées, considérons un premier exemple : 'la marche que Paul a faite jusqu'à la gare ce matin'. Cet événement concerne la vie de multiples processus, dont des processus de marche de Paul (en utilisant le pluriel, nous imaginons que Paul a pu s'arrêter puis reprendre sa marche, autrement dit que plusieurs processus de marche ont successivement existé), et la vie de multiples objets, dont Paul lui-même. Plus généralement, nous considérons qu'un *événement* est quelque chose :

e_i) existant pour un sujet à des temps ;
e_ii) jouissant d'une méréologie *dérivée* et *robuste*
e_iii) pouvant *survenir*, mais sans être répété

Un événement est avant tout un objet de pensée, une entité psychologique, servant à des sujets à se représenter l'histoire d'un monde (physique ou non). Le terme « histoire » évoque

généralement une histoire *passée*. Je pense ainsi à 'ma dernière chute de vélo' et à d'autres événements survenus dans le passé et faisant partie de notre imaginaire collectif : 'l'assassinat de César par Brutus', 'le naufrage du Titanic'[4]. Mais l'histoire peut être *présente* : je pense à l'événement 'l'écriture de cet article' au moment même où j'écris l'article. Elle peut également être *future* : je pense au 'déplacement en voiture que je dois effectuer demain'.

Comme le montrent ces exemples, les événements existent dans des espaces-temps distincts de ceux dans lesquels existent les substances dont ils relatent l'histoire. Ce sont des entités *abstraites* endurant dans l'espace cognitif d'un sujet (e_i). Cette caractérisation les oppose aux événements concrets de Davidson (1969, 1970) et entraîne de nouvelles propriétés.

Les deux espaces-temps étant découplés, il convient de noter que les événements jouissent d'une méréologie *dérivée* (e_ii). Ainsi, dire que l'événement 'l'enfance de Paul' « fait partie » de l'événement 'la vie de Paul' ne signifie pas que les deux événements partagent une même région spatio-temporelle dans l'esprit d'un sujet les pensant. Il faut plutôt entendre que les faits concernés par le premier événement font partie (dans un sens ensembliste) des faits concernés par le second événement. À ce propos, on peut se demander dans quelle mesure les faits concernés par un événement sont déterminés ?

Sur ce point, on peut noter que des événements « simples » correspondent à des épisodes de vie d'un seul objet ou d'un seul processus bornés dans un espace-temps déterminé, comme dans le cas de 'la chute de Paul à l'instant'. Cet exemple illustre la différence conceptuelle existant entre le processus et l'événement, ce dernier étant obtenu en fixant des limites temporelles à un processus et à le considérer dans la durée, comme le soutiennent Galton et Mizoguchi (2009, p. 75) :

> *We maintain, on the contrary, that so far from being a mark of short duration, boundedness is a precondition for the assignment of any definite duration: processes endure, but only once we have assigned bounds to them can we speak of duration, and the act of assigning bounds means that we have switched our attention from the process to an event.*

Par contre, comme nous l'avons déjà fait remarquer, certains événements, tel 'la marche que Paul à faite ce matin jusqu'à la gare', sont plus complexes en s'accommodant de faits très différents. Cet événement ne tient pas compte de l'itinéraire suivi par Paul pour se rendre à la gare. On peut parler ici de *robustesse* méréologique des événements, par opposition à l'*essentialisme* méréologique qui leur est souvent attribué (à savoir, le fait qu'un événement soit essentiellement déterminé par ses parties). Comme le note Achille Varzi (2002), en considérant des événements correspondant à des épisodes de vie de larges systèmes économiques et politiques tels 'la révolution industrielle' ou 'la seconde guerre mondiale', les événements comportent une indétermination intrinsèque.

Poursuivons notre caractérisation des événements en évoquant une propriété qui les qualifie ordinairement d'« occurrents », à savoir le fait qu'ils puissent *survenir* (e_iii). Le même qualificatif est également ordinairement attribué aux processus mais, dans leur cas, nous nous sommes contentés d'évoquer leur existence. Pour les événements, les choses se présentent différemment et ceci est dû au fait de les identifier à des objets de pensée. Ceci conduit en effet à distinguer leur *existence* de leur *survenue*. La définition générale que nous retenons de cette propriété est la suivante :

> Soit *e* un événement existant pour le sujet *s* au temps *t* ; l'événement *e* 'survient' à un temps *t'* ssi les faits dont *e* relate l'histoire existent au temps *t'*.

[4] Dans cet article, pour des raisons de place et compte tenu de notre propos, nous ne développerons pas la dimension sociale des événements et nous contenterons donc de considérer des événements pensés par un sujet unique, qui plus est humain.

La propriété de *survenue* des événements peut être considérée comme analogue à la propriété de *véracité* des propositions : l'existence de faits conditionne la survenue de l'événement, comme elle conditionne la vérité d'une proposition. L'existence de faits de chute de Paul à des instants consécutifs $t_1$ et $t_2$ - <Paul,Tombe,$t_1$>$_F$ et <Paul,Tombe,$t_2$>$_F$ – correspond à la condition de survenue de l'événement <Paul,Chute,À l'instant>$_E$ (pour peu que l'événement soit pensé et que le temps pensé 'À l'instant' corresponde à l'intervalle [$t_1$,$t_2$]). La relation d'ordre entre *t* et *t'* dans la définition détermine le fait que l'histoire soit passée, présente ou future. Dans le cas où l'événement survient au moment même où il est pensé, cela donne la possibilité au sujet de jouer le rôle d'agent en agissant sur les faits réalisant l'événement[5].

Tout dernier point, que nous mentionnerons sans le justifier. Nous considérons qu'il existe des événements particuliers non répétés (ex : 'la chute de Paul de ce matin') et des types d'événements répétables (ex : 'la marche matinale de Paul jusqu'à la gare'). Intuitivement, la singularité est liée aux faits concernés : un épisode de vie d'un processus *vs* une classe d'épisodes de vie d'une classe de processus (ce critère serait à généraliser à des événements ne se référant pas à des processus).

### 2.5 En résumé

En conclusion de cette section, résumons avec la figure 2 le chemin parcouru depuis la distinction entre entités *expérientielles* et *historiques* de Galton (2008) qui nous a servi de point de départ.

Le monde tel qu'il se présente à nous est peuplé en premier lieu de substances – des objets et des processus physiques. Ces substances, dont on peut constater l'existence à tout niveau d'abstraction, dépendent mutuellement l'une de l'autre. Une raison toutefois de les distinguer tient au fait qu'elles portent des propriétés différentes (Hacker, 1982b) : un processus physique n'a pas plus de couleur ou de volume qu'un objet n'a de rapidité ou d'amplitude. Toujours dans le monde physique, les substances, tout en portant temporairement des propriétés, jouissent d'une vie correspondant à des faits physiques. Parmi ces faits physiques figurent les perpétuations de processus jouant un rôle important dans la dynamique du monde physique.

Pour basculer cette fois dans le réel construit, des sujets cognitifs pensent le monde et, parmi ces objets de pensée, figurent les événements. Positionnés à côté des propositions, nous considérons que ces entités, représentant le monde, jouent des rôles cognitifs distincts : si la notion de *vérité* caractérise les propositions, celle de *survenue* caractérise les événements. Par ailleurs, du côté des objets de pensée figurent les faits sociaux rendant compte notamment de la vie sociale des substances.

Dans cette section, nous avons présenté un cadre ontologique général en nous attachant à souligner sa cohérence globale. On notera toutefois que nous avons laissé momentanément de côté l'analyse des objets et processus non-physiques, indispensables pour rendre compte de phénomènes comme 'la dépréciation du dollar par rapport à l'euro'. Ceci signifie que notre inventaire ontologique reste à compléter, notamment en vue de proposer une ontologie qui puisse être utilisée en Ontologie Appliquée (à l'instar de BFO et DOLCE). En attendant, en guise de conclusion de cet article, nous discutons la question figurant dans le titre, à savoir : les distinctions ontologiques que nous venons de voir offrent-elles une alternative à l'opposition 'continuant' vs 'occurrent' ?

[5] Dans (Kassel, 2017), nous détaillons de telles situations en les caractérisant en termes de couplages temporels et causaux entre processus et événements.

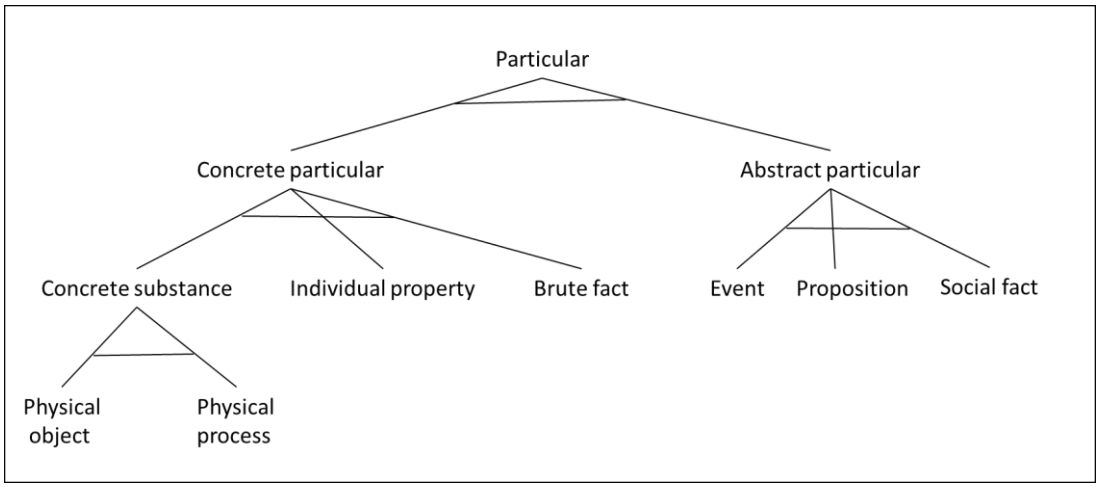



FIGURE 2 – Vue d'ensemble de notre cadre ontologique

## 3 Discussion

Le cadre ontologique que nous venons d'esquisser remet en cause l'opposition 'continuants' vs 'occurrents' adoptée classiquement en Ontologie Appliquée, notamment par BFO (Grenon & Smith, 2004) et DOLCE (Masolo *et al.*, 2003). Cette opposition, rappelons-le, se fonde sur deux doctrines philosophiques caractérisant deux modes distincts d'existence et de persistance pour les entités peuplant le monde : l'*endurantisme* (le fait pour une entité d'exister dans son entièreté à tout moment de son existence et de pouvoir changer dans le temps) et le *perdurantisme* (le fait pour une entité de n'exister que par ses parties à tout moment et de ne pas pouvoir changer dans le temps). Fondamentalement, la remise en cause porte sur deux points :

1) le fait d'admettre dans la classe des continuants une nouvelle catégorie d'entités rendant compte de la dynamique du monde – les *processus physiques* – avec pour conséquence éventuelle de revoir la notion de *continuant* ;

2) le fait de devoir réviser radicalement la classe des *occurrents* pour accueillir les *événements* conçus comme des objets de pensée, mais également de devoir faire une place aux faits de perpétuation de processus.

### 3.1 De la notion de *continuant*

La figure type de continuant est celle de l'objet physique matériel, habituellement conçu comme une entité 3D. Cette conception, dont nous rendons compte avec les propriétés (o_i) (o_ii)(o_iii), suppose de distinguer deux notions (Fine, 2006) : l'*existence* de l'objet (souvent dénotée également par le terme « présence ») et son *extension* (ou « localisation ») dans l'espace. La première possède un caractère de tout ou rien – un objet existe, ou non – tandis que la seconde possède un caractère de partialité : si l'objet occupe *totalement* une région spatiale, il n'occupe que *partiellement* chaque partie de cette région spatiale. En ce sens, nous pouvons dire que l'objet existe (ou qu'il est présent) « dans sa totalité » ou « dans son entièreté » à des temps, mais qu'il n'existe (ou qu'il n'est présent) qu'en partie dans certaines régions spatiales à des temps. En revanche, l'objet physique ne possède pas d'extension temporelle et cette conception est à opposer à celle de l'entité 4D proposée par certains métaphysiciens (ex : (Sider, 1997)) considérant qu'un objet physique est étendu dans le temps

de la même façon qu'il est étendu dans l'espace. L'argument que nous mettons en avant pour défendre la conception 3D de l'objet physique est qu'il s'agit d'une conception intuitive de sens commun alors que la conception 4D, d'une part, est largement contre-intuitive et, d'autre part, n'est étayée par aucune donnée de psychologie ou quelconque matériau linguistique. Une conception 4D ne peut donc fonder une ontologie descriptive telle que visée.

Venons-en maintenant aux processus. En les caractérisant au moyen des propriétés (p_i)(p_ii)(p_iii), nous les avons assimilés à des continuants « occurrents » ou « dynamiques », pour reprendre les termes et la conception de Stout (2016). Une telle thèse a été critiquée par Steward (2015), arguant du fait que les processus demeurent des occurrents possédant une extension temporelle, ce qui contredit (p_i) :

> *Because processes have temporal parts and temporal extension, they do not exist in their entirety at each moment of their existence, in the way that substances do. A process is an essentially unfolding entity, which is what secures its right to be thought of as an occurrent.*

En conséquence, toujours selon Steward, il est faux de considérer que les processus aient leurs propriétés *à* des temps (p_ii). Il convient plutôt de considérer qu'ils ont leurs propriétés *entre* des temps (2015) (nous soulignons ce qui apparaît en italique dans l'article) :

> *[Processes] have their properties primarily between times, and this implies that they share some of the characteristics continuants are generally thought to have […] To say that the dripping of a tap was persistent at t1 is a way of saying that t1 was a moment which falls within a period of time over which the dripping was persistent.*

Un élément important du débat nous paraît être la définition en extension de la classe des processus. On notera à ce propos que l'exemple analysé par Steward d'un robinet qui fuit – le seul exemple de processus, du reste, traité dans son (2015) – ne correspond pas à notre conception du processus physique. Sémantiquement, le terme 'fuite' évoque un flux d'eau s'échappant anormalement d'un robinet. Supposons que la fuite soit intermittente : chaque flux correspond, selon notre conception, à un processus énacté par une goutte d'eau (ou une quantité d'eau plus importante) ; on peut ajouter que chaque échappement de flux correspond à un processus énacté, cette fois par le robinet ; par contre, la fuite en elle-même, en tant qu'épisode correspondant à l'existence de ces différents processus, est au mieux un événement correspondant à l'histoire du robinet et de ces processus (le terme « au mieux » soulignant la condition qu'un sujet doive penser cette histoire pour qu'elle existe) ; par ailleurs, l'existence d'un événement n'implique pas qu'un processus le constituant existe tout au long de l'événement (surtout si l'événement est qualifié d'« intermittent »).

Mais l'élément essentiel des discussions porte sur la notion même d'*existence*. Nous avons vu que la locution « existe dans son entièreté », lorsqu'elle se rapporte aux objets physiques, signifie « l'objet existe et il est étendu totalement sur une région spatiale ». La question reste donc celle du sens à donner à la proposition-propriété (o_i) « l'objet existe à des temps ». Précisons que nous ne considérons pas l'existence comme une propriété physique. Rappelons que nous avons adopté la conception Armstrongienne des propriétés comme des *manières d'être* d'objets et non l'*être* même. Inutile donc de chercher du côté des faits pour des vérifacteurs de telles propositions existentielles. Suivant Peter Simons (2000), nous considérons qu'un objet dépend existentiellement de processus. Pour un être vivant, la biologie nous renseigne sur la nature de ces processus (2000, p. 70) :

> *For a human being or other animal the relevant processes are those which are vital to it, which are a (probably not exactly delimited) collection of occurrents in its life, involving respiration, blood transport, nutrient breakdown and the chemical reactions within the cells.*

Pour des objets inertes comme une pierre, ce sont la physique et la chimie qui nous renseignent sur la nature de ces processus (2000, p. 71) :

> *What keeps the rock in existence? If we look more closely into the physics and chemistry of it we find it much less boring. The widespread cohesion of the crystals in the rock which hold it together as a mechanically unified mass depend on chemical bonds among atoms, and these depend on sharing and exchange of electrons and on the continued existence of the many particles in the rock […] The rock is in fact teeming with occurrents, vast numbers of them, and they are vital to it: if they stopped, it would cease to exist.*

Nous retrouvons notre conception d'une dépendance mutuelle entre objets et processus telle qu'exprimée par les propriétés (o_iv) et (p_iv). Notamment, pour les processus, leur existence dépend d'objets qui, à leur tour, dépendent existentiellement de processus, *etc*. Nous en concluons une proximité d'existence entre processus et objets, suffisante pour classer les processus dans la catégorie des continuants. Par ailleurs, nous interprétons ces dépendances mutuelles se retrouvant à différents niveaux d'abstraction (une position défendue par Galton et Mizoguchi (2009)) comme le fait qu'aucune de ces deux primitives ne soit métaphysiquement prioritaire sur l'autre.

### *3.2* De la notion d'*occurrent*

Pour revenir sur la classe des entités *occurrentes*, c'est-à-dire des entités dont on dit qu'elles « surviennent », ayant rangé les processus du côté des continuants concrets, nous avons identifié deux autres catégories d'occurrents : les *perpétuations* de processus (ex : la poussée que j'exerce sur la porte *perpétue* son ouverture), dont nous avons fait des particuliers concrets, et les *événements*, positionnés du côté des entités abstraites et classiquement considérés dans la littérature comme des perdurants.

Les faits de perpétuation de processus ne figurent ordinairement pas dans les inventaires d'entités occurrentes. La raison que nous avançons est que la conception prédominante des processus, suivant les travaux de Mourelatos (1978), est d'admettre des « macro » processus comme 'ouvrir la porte', 'écrire une lettre', voire même 'construire une maison'. Ces macro-processus se retrouvent dans la doctrine prédominante de la constitution des événements par des processus, conduisant à considérer que des événements comme 'l'ouverture d'une porte' ou 'la construction d'une maison' sont constitués a minima d'UN processus présent toute la durée de l'événement. Notre conception du processus physique met à mal cette doctrine et nous conduit au contraire à identifier plusieurs processus physiques se propageant les uns les autres : une ouverture manuelle d'une porte suppose ainsi l'existence d'un mouvement de la main perpétuant un mouvement de la porte. Sans ces faits de perpétuations, nous serions en peine de rendre compte de ces phénomènes dynamiques.

Côté événements, rompant avec la doctrine de la constitution des événements par les processus, nous leur avons en effet accordé un statut d'entités abstraites, reprenant une proposition faite par Kathleen Gill en conclusion de son (1993), seule possibilité selon elle pour préserver les deux catégories « processus » et « événements » (1993) :

> *Just as physical objects apparently form a metaphysical subcategory of objects distinct from, e.g., numbers and spiritual objects, so a more appropriate starting point for developing a metaphysical subcategorization of occurrences would be to distinguish physical occurrences from, e.g., negative, merely possible or perhaps spiritual occurrences.*

Comme en témoigne le terme « occurrence spirituelle » (nous pourrions évoquer également le terme « objet non existant »), le seul fait de ranger les événements du côté des entités

abstraites les apparente aussitôt, pour la plupart des métaphysiciens, à des citoyens de « seconde classe ». Nous le voyons ainsi dans le projet de Joseph Mélia (2000) de proposer une ontologie reposant uniquement sur des continuants : selon Mélia, si des discours courants portent sur des événements (ex : une fête, un match de football) semblant indiquer une existence, ces derniers n'existent au mieux qu'en tant qu'entités linguistiques mais n'intéressent pas vraiment les métaphysiciens. Ceci explique que l'ontologie des événements reste largement à établir.

Dans cet article, nous avons avancé des pistes en montrant que le perdurantisme ne s'appliquait pas aux événements. La raison essentielle est que les événements sont caractérisés par une méréologie indirecte : s'ils sont étendus dans le cerveau du sujet qui les pense, ce qui compte pour des sujets est qu'ils relatent des vies occupant des régions spatiotemporelles. Du fait qu'ils relèvent de l'histoire de la vie des substances, leur existence (pour un sujet) ne dépend pas de celle des vies en question. Ce sont des continuants psychologiques (voire sociaux) permettant à des sujets d'embrasser des épisodes de vie pouvant *survenir*, cette survenue dépendant de l'accumulation de faits (on retrouve ici l'intuition du perdurantisme selon laquelle un événement comme un match de football ne peut survenir qu'à la condition qu'un ensemble de faits existent consécutivement).

Cette caractérisation interroge sur le (ou les) rôle(s) cognitif(s) joué(s) par les événements, notamment par rapport aux propositions. Dans (Kassel, 2018) nous avons proposé d'identifier notamment les événements à des contenus d'intentions, en leur faisant jouer un rôle dans les mécanismes de spécification et de contrôle d'action. De quoi relativiser leur statut de « citoyen de seconde classe ».

## Références


ARMSTRONG D.M. (1997). *A world of states of Affairs*. Cambridge University Press.

BERGSON H. (1934). *La pensée et le mouvement ; essais et conférences*. Félix Alcan, Paris.

BORGO S., FRANSSEN M., GARBACZ P., KITAMURA Y., MIZOGUCHI R., VERMAAS P.E. (2014). Technical Artifacts: An integrated perspective. *Applied Ontology*, 9(3-4), 217-35.

CHISHOLM R. (1970). Events and Propositions. *Noûs*, 4(1), 15-24.

CROWTHER T. (2011). The Matter of Events. *The Review of Metaphysics*, 65(1), 3-39.

CROWTHER T. (2018). Processes as Continuants and Process as Stuff. In R STOUT (ed.), *Process, Action, and Experience*, Oxford University Press, pp. 58-81.

DAVIDSON D. (1969). The Individuation of Events. In N. RESCHER (ed.), *Essays in Honor of Carl G. Hempel* (pp. 216-34), Dordrecht: D. Reidel.

DAVIDSON D. (1970). Events as Particulars. *Noûs*, 4(1), 25-32.

DRETSKE F. (1988). *Explaining behavior: Reasons in a world of causes*. Cambridge, MA: MIT Press.

FERRARIS M. (2014). *Manifeste du nouveau réalisme*. Hermann : Paris ; traduction française de M. Flusin et A. Robert de l'ouvrage (2012) : *Manifesto del nuovo realismo*.

FINE K. (1982). First-Order Modal Theories III – Facts. *Synthèse*, 53, 43-122.

FINE K. (2006). In Defense of Three-Dimensionalism. *The Journal of Philosophy*, 3(12), 699-714.

GALTON A. (2006) On What Goes On: The ontology of processes and events. In R. FERRARIO & W. KUHN (eds.), proceedings of the Fourth International Conference on Formal Ontology in Information Systems (FOIS2006), pp. 4-11.

GALTON A. (2008). Experience and History: Processes and their Relation to Events. *Journal of Logic and Computation*, 18(3), 323-40.

GALTON A. (2012). States, Processes and Events, and the Ontology of Causal Relations. In M. DONNELLY & G. GUIZZARDI (eds.), proceedings of the 7th international conference on Formal Ontology in Information Systems (pp. 279-92), IOS Press.

GALTON A. & MIZOGUCHI R. (2009). The water falls but the waterfall does not fall: New perspectives on objects, proceses and events. *Applied Ontology*, 4(2), 71-107.

GILL K. (1993). On the Metaphysical Distinction Between Processes and Events. *Canadian Journal of Philosophy*, 23(3), 365-84.

GRENON P. & SMITH B. (2004). SNAP and SPAN: Towards dynamic spatial ontology. *Spatial Cognition and Computation*, vol. 4, n° 1, p. 69-104.

HACKER P.M.S. (1982a). Events, Ontology and Grammar. *Philosophy*, 57, 477-86.

HACKER P.M.S. (1982b). Events and Object in Space and Time. *Mind*, 91, 1-19.

HORGAN T. (1978). The Case Against Events. *Philosophical Review*, 87(1), 28-47.

HORNSBY J. (2012). Actions and activity. *Philosophical Issues*, 22(1), 233-45.

IACONA A. (2003). Are there propositions? *Erkenntnis*, 58(3), 325-51.

KASSEL G. (2010). A formal ontology of artefacts. *Applied Ontology*, 5(3-4), 223-46.

KASSEL G. (2017). Processus, événements et couplages temporels et causaux. *Revue d'Intelligence Artificielle*, 31(6), p. 649-679.

KASSEL G. (2018). Ontologie de l'action et formes logiques des phrases d'action : de nouvelles perspectives. In T. DE LIMA et S. DOUTRE (eds.), Actes des 12èmes Journées d'Intelligence Artificielle Fondamentale, Amiens, 13-15 juin.

MARGOLIS E. & LAURENCE S. (Eds.)(1999). *Concepts: Core readings*. MIT Press.

MASOLO C. (2010). Understanding Ontological Levels. In F LIN & U SATTLER (eds.), proceedings of the 12th International Conference on the Principle of Knowledge Representation and Reasoning (KR 2010), pp. 258-268.

MASOLO C., BORGO S., GANGEMI A., GUARINO N., OLTRAMARI A. & SCHNEIDER L. (2003). The WonderWeb Library of Foundational Ontologies and the DOLCE ontology. WonderWeb Deliverable D18, Final Report, vr. 1.0.

MÉLIA J. (2000). Continuants and Occurrents. *Proceedings of the Aristotelian Society, Supplementary Volumes*, 4, 77-92.

MOURELATOS A.P.D. (1978). Events, Processes, and States. *Linguistics and Philosophy*, 2(3), 415-434.

SEARLE J.R. (1995). *The Construction of Social Reality*. NewYork: Free Press.

SEARLE J.R. (2010). *Making The Social World: The Structure of Human Civilization*. Oxford University Press.

SEIBT J. (2008). *Beyond Endurance and Perdurance: Recurrent Dynamics*. In C. KANZIAN (ed.), Persistence, Frankfurt: Ontos Verlag, pp. 133-164.

SIDER T. (2001). *Four-dimensionalism: An Ontology of Persistence and Time*. Oxford: Oxford University Press.

SIMONS P. (2000). Continuants and Occurrents. *Proceedings of the Aristotelian Society, Supplementary Volumes*, 4, 59-75.

STEWARD H. (2012). Actions as Processes. *Philosophical Perspectives*, 26(1), 373-88.

STEWARD H. (2013). Processes, Continuants, and Individuals. *Mind*, 122(487), 781-812.

STEWARD H. (2015). What is a continuant? In Proceedings of the Aristotelian Society, Supplementary (pp. 109-123), Volume: LXXXIX.

STOUT R. (1997). Processes. *Philosophy*, 72(279), 19-27.

STOUT R. (2003). The life of a process. In G. DEBROCK (ed.), Process Pragmatism: Essays on a Quiet Philosophical Revolution (pp. 145-57), Rodopi.

STOUT R. (2016). The category of occurrent continuants. *Mind*, 125(497), 41-62.

STRAWSON P. (1959). *Individuals. An Essay in Descriptive Metaphysics*. Methuen, London.

THOMASSON A.L. (2003). Foundations for a social ontology. *Protosociology*, vol. 18, 269-90.

VARZI C.A. (2002). Events, Truth and Indeterminacy. *The Dialogue*, 2, 241-264.

WILSON, N. (1974). Facts, Events, and Their Conditions. *Philosophical Studies*, XXV, 303-321.

# oogo : Ontologie des Outils utiles à la Gestion d'Ontologies

Sylvie Despres

Université Paris 13, Sorbonne Paris Cité, LIMICS, (U1142), INSERM, Sorbonne Universités, UPMC Université Paris 6,
74 rue Marcel Cachin F-93017 Bobigny cedex, France,
sylvie.despres@univ-paris13.fr

**Résumé** : Cet article présente, oogo, une ontologie ayant pour objectif la description des outils utiles à la gestion d'ontologies. A l'origine, ce travail répondait à une demande qui était de présenter à des chercheurs et/ou des industriels des outils utilisables pour élaborer une ontologie de domaine et si nécessaire leur permettre de les sélectionner en fonction de leurs besoins. Une ontologie nous a semblé constituer la forme la plus adaptée pour délivrer cette information. Le modèle conceptuel la structurant est fondé sur le cycle classique de construction d'ontologies commun à toutes méthodologies. Les activités afférentes à la construction d'ontologies servent de guide à la présentation des fonctionnalités les plus utiles de ces outils. Les scénarios d'usage sont construits à partir des besoins des acteurs intervenant autour d'une ontologie. Un état de l'art relatif à chaque catégorie d'outils sous tend la construction du modèle conceptuel de oogo.

**Mots-clés** : oogo, ontologie, ingénierie des ontologies, outils pour l'ingénierie ontologique.

## 1 Introduction

Dans cet article, nous décrivons, oogo, une ontologie ayant pour objectif la description des outils utiles à la gestion d'ontologies. A l'origine, ce travail répondait à une demande des organisateurs de la journée PDIA2017[1] qui était de présenter à des chercheurs et/ou des industriels des outils utilisables pour élaborer une ontologie de domaine et si nécessaire leur permettre de les sélectionner en fonction de leurs besoins.

La première question qui s'est posée pour répondre à cette requête a été de trouver la forme la plus pertinente pour renseigner ces acteurs. Il existe en effet de nombreux articles décrivant les différents outils et de nombreuses comparaisons les concernant. Ces dernières sont difficiles à exploiter car elles sont statiques, dédiées à des catégories d'outils particulières et les critères de comparaison utilisés peuvent varier d'un article à l'autre. Dans ce contexte, l'utilisation d'une ontologie nous a semblé la forme la mieux adaptée pour donner accès à l'ensemble des caractéristiques des outils actuellement disponibles. Elle permet entre autres de répondre aux questions les plus fréquemment posées lors de la sélection des outils intervenant dans la gestion d'une telle ressource. Elle est destinée à être publiée pour être réutilisée et enrichie par la communauté d'ingénierie des ontologies. Les utilisateurs concernés par oogo sont des ingénieurs de la connaissance ou des chargés de mission en « information technology » en charge de la construction d'ontologies et s'interrogeant sur les fonctionalités des outils existants.

Le modèle conceptuel de oogo est fondé sur les activités intervenant dans le cycle classique de construction d'ontologies (scénario 1 de la méthodologie Neon) commun à toutes méthodologies de construction (Suárez-Figueroa et al., 2015). Les activités afférentes à la

[1] https://afia.asso.fr/wp-content/uploads/2018/01/cr-PDIA-2017-vf.pdf

construction d'ontologies servent de guide à la présentation des fonctionnalités les plus utilisées de ces outils (Suárez-Figueroa, Gomez Perez, 2008). Le périmètre de l'ontologie est établi en adoptant la méthodologie de (Uschold et Gruninger, 1996). Les scénarios d'usage identifiés sont issus de cas réels rencontrés lors de la construction d'ontologies. A partir de ces scénarios, une liste non exhaustive de questions de compétence, auxquelles oogo doit être en mesure de répondre, a été élaborée à partir des questions posées par les acteurs devant construire une ontologie. Un état de l'art relatif à chaque outil aide à la construction du modèle conceptuel de oogo.

Dans la suite de l'article, la section 2 décrit le modèle adopté pour le concept Outil à partir duquel les modèles des outils présentés dans la section 3 sont décrits. Puis nous concluons dans la section 4 sur l'intérêt et l'usage d'une telle ontologie.

## 2 Modèle adopté pour le concept Outil

Nous envisageons l'outil comme un moyen permettant d'obtenir un résultat correspondant à la mise en œuvre d'un service. Nous proposons un modèle pour le concept d'outil dédié à la construction de ressources du web sémantique que nous spécialiserons en fonction du type d'outil étudié.

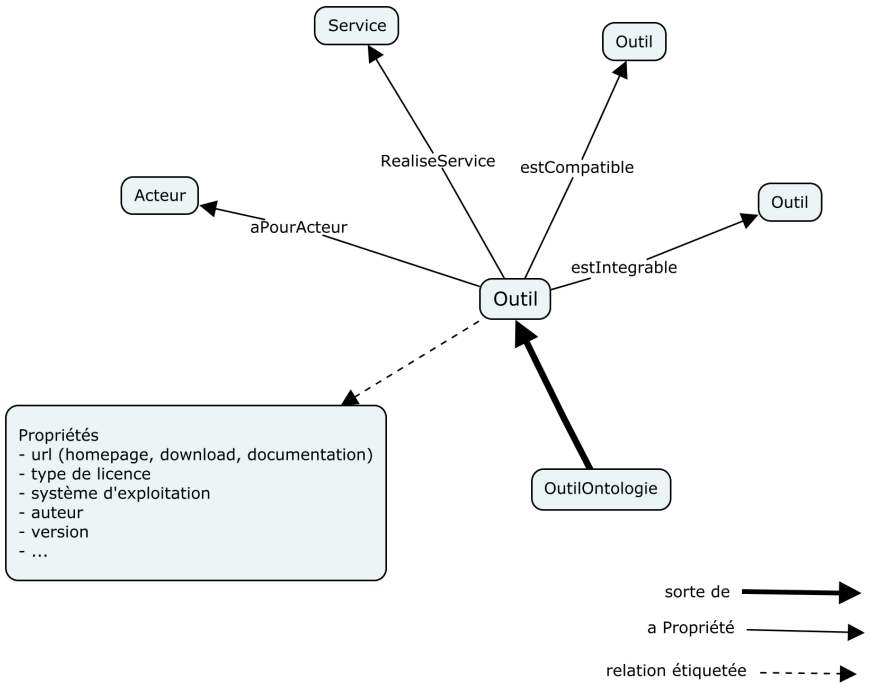



FIGURE 1 – Modèle du concept Outil

En utilisant plusieurs vocabulaires standards, un outil est identifié par une URI correspondant à sa homePage, une URI de documentation et une URI de téléchargement. Il peut être caractérisé par des mots-clés. Il a été conçu par un auteur ou une équipe d'auteurs et développé par une organisation. Il est exécutable sous un SE. Il est identifié par un type de licence. Il est mis à disposition selon différentes modalités. Il fournit un type de service. Il est interopérable avec d'autres outils. Il dispose d'un format de stockage pour les fichiers contenant les ressources. Il permet l'import et l'export de ressources. Il est capable de gérer les sauvegardes avec ou sans gestion de version.

Plusieurs vocabulaires sont réutilisés pour la construction du modèle concept Outil : dcterms, doap, dvia, educore, foaf, org.

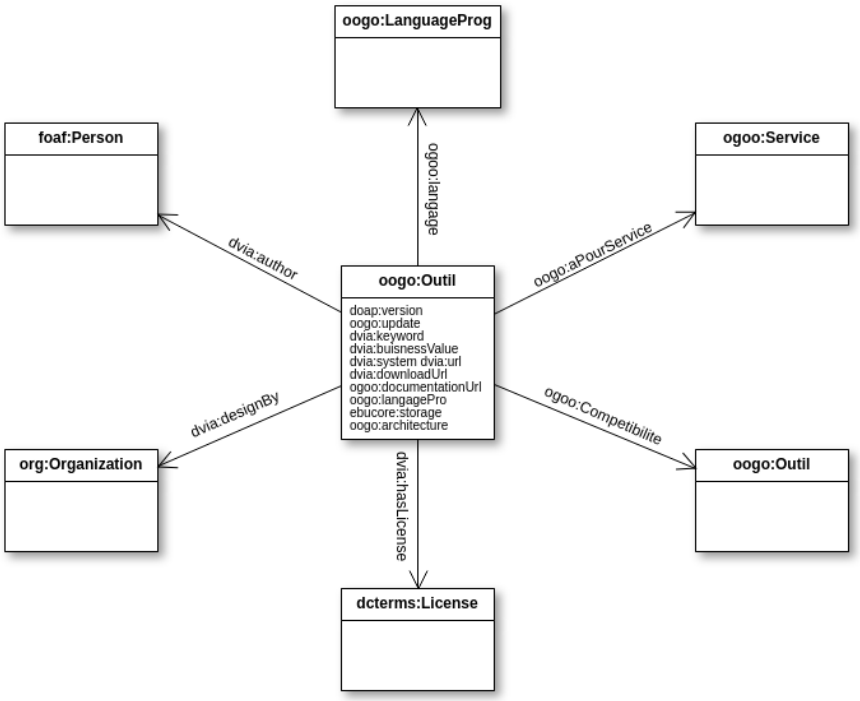



FIGURE 2 – Aperçu des vocabulaires réutilisés

## 3 Démarche de construction de l'ontologie oogo

Nous nous intéressons aux outils dédiés à la construction d'ontologie définie comme « une spécification formelle explicite d'une conceptualisation partagée » (Studer et al., 1998). Le langage de représentation considéré pour la formalisation de ces ontologies est OWL/OWL2 standard du W3C.

Le cycle classique de développement d'une ontologie est constitué des étapes de spécification, planification, conceptualisation, formalisation, implémentation (Suárez-Figueroa et al., cf. *supra*). Au cours de ce cycle, les activités mises en œuvre (Suárez-Figueroa, Gomez Perez, cf. *supra*) comportent les activités de gestion, orientées développement (pré-développement, développement, post-développement) et de support à la construction. Dans oogo, nous décrivons une partie des outils supportant les activités de support et de développement.

Le cadre méthodologique dans lequel nous nous situons est celui du projet NeON (http://www.neon-project.org/). Nous adoptons le cycle classique de construction d'une ontologie (spécification, planification, conceptualisation, formalisation, implémentation) correspondant au scénario 1 de la méthodologie Neon.

La spécification des besoins auxquels doit répondre l'ontologie a été décrite dans l'introduction. La sélection des outils présentés reposent sur des critères qui sont relatifs à la disponiblité pour une installation, leur maintenance, leur actualité et leur formalisation en OWL. Elle est non exhaustive et pourra être étendue au fil de l'utilisation de l'ontologie. Une

partie a été collectée *via* le site du w3c[2] et de AI[3] [(3)] et les articles synthétisant les différents outils existants. Ces outils ont servi à l'élaboration des modèles leur correspondant et sont décrits dans oogo.

Nous avons identifié, pour chacun d'entre eux, les scénarios d'usage issus de l'expérience acquise au cours des différents projets de recherche menés au LIMICS et les questions de compétences qui leur sont associées, c'est-à-dire les questions auxquelles notre ontologie doit être en mesure de répondre. La construction des modèles de connaissances qui leur sont associés, est décrite dans la suite de ce paragraphe.

## 3.1 Activités de support

Les activités de support que nous considérons sont la modélisation des connaissances, la réutilisation, l'évaluation et la validation. Les outils pouvant supporter ces activités sont les outils de recherche d'ontologies (moteurs de recherche et bibliothèques d'ontologies) et les outils de validation.

### 3.1.1 Outils servant à la modélisation

Des outils de modélisation graphique semi-formelle peuvent être utilisés dans la phase de conceptualisation. Ils permettent de construire le modèle conceptuel du domaine qui sera ensuite formalisé avec l'éditeur d'ontologie. Leur utilisation intervient au cours de la phase d'acquisition de connaissances qui est réalisée de manière incrémentale.

#### *3.1.1.1 Les scénarios*

Scénario 1 : L'utilisateur cherche à cerner le périmètre de l'ontologie à construire. Il doit choisir une représentation sous forme de carte cognitive.

Scénario 2 : L'utilisateur cherche à représenter graphiquement le modèle du domaine en partant d'un concept central. oogo doit décrire les outils de modélisation avec leurs propriétés.

#### *3.1.1.2 Les questions de compétence*

Question 1 : J'ai identifié un concept central et je cherche à décrire les entités connexes en indiquant les relations qui les lient. Quel outil me permet de le faire ?

Question 2 : Comment déterminer la granularité de la représentation ?

Question 3 : Est-il possible de traduire automatiquement les cartes obtenues au format OWL ?

#### *3.1.1.3 Les outils de modélisation sélectionnés*

Nous avons sélectionné les outils Yed Graph Editor[4], XMind[5] et CmapsTools[6]. Ces trois outils disposent de versions exploitables sousWindows, OsX et Linux. Nous ne considérons pas dans cet article les métamodèles de représentation de OWL dans le formalisme UML et le langage GOWL de modélisation graphique, polymorphique et typé pour la construction d'ontologie (Héon et *al*., 2016).

Yed Graph Editor est un éditeur de graphes de bas niveau, open source et multiplateforme, personnalisable par programme. Il est souple dans les options de présentation disponibles et possède plusieurs fonctionnalités de disposition automatique des éléments du graphe. Il permet d'exporter les graphes au format graphml, SVG, png, emf, eps, etc.

---

[2] https://www.w3.org/wiki/

[3] http://www.mkbergman.com/904/listing-of-185-ontology-building-tools/

[4] http://www.yworks.com/yed

[5] http://www.xmind.net/

[6] https://cmap.ihmc.us/

XMind est un outil de carte heuristique qui permet également la construction de carte conceptuelle. Il permet de créer, gérer et partager des cartes conceptuelles et de les exporter au formats txt, html, image, SVG. XMind propose une version open source et une version payante. Dans la seconde version, il est possible d'exporter au format csv. Cette fonctionnalité peut ensuite être exploitée par d'autres outils de création semi-automatisée de l'ontologie.

CmapTools est un outil qui permet de créer et partager facilement des cartes conceptuelles. Il est utilisable en ligne *via* un simple navigateur ou téléchargeable sous la forme d'un exécutable. Il dispose d'options de mise en page automatique et permet d'exporter au format image, PDF, etc. Il permet de fusionner des nœuds identiques, de créer des nœuds imbriqués de valider des liens, de réparer automatiquement les liens rompus et d'annoter les nœuds. Il permet également la comparaison des cartes qui est une fonctionnalité utile pour la construction collaborative.

#### *3.1.1.4 Le modèle pour les outils de modélisation*

Le modèle du concept Outil de modélisation prend en compte les fonctionnalités qui leur correspondent. Nous avons en particulier retenu le mode de mise en forme, les formats d'export, leur licence, les options de partage, le mode d'exploitation (en ligne ou téléchargeable). Les notions de cartes heuristique et conceptuelle sont décrites.

### 3.1.2 Outils de recherche

#### *3.1.2.1 Les scénarios*

Scénario 1 : L'utilisateur cherche à réutiliser des ressources ontologiques pour construire sa propre ressource. Il s'interroge sur les outils lui permettant de réaliser cette recherche et le type de recherche qu'il peut effectuer. oogo doit représenter les outils avec les types de recherche associés.

Scénario 2 : L'utilisateur s'interroge sur le type de recherche utilisé par l'outil pour retrouver une ressource. oogo doit décrire les différentes modalités de recherche associées aux outils.

Scénario 3 : L'utilisateur a réalisé une recherche qui lui renvoie une liste de ressources, il cherche une explication du classement des ressources retournées. oogo doit préciser les métriques utilisée pour le classement (ranking) des ressources.

#### *3.1.2.2 Les questions de compétence*

Question 1 : Quels outils de recherche existent pour m'aider à trouver une ontologie de domaine ? (scénario 1)

Question 2 : Existe-t-il des vocabulaires dont je pourrais réutiliser les termes pour désigner les concepts ou les relations de mon ontologie ? (scénario 1)

Question 4 : Quelles sont les modalités de recherche mises en œuvre pour trouver une ressource ? (scénario 2)

Question 5 : Comment s'explique la classification des ressources proposées par l'outil de recherche ? (scénario 3)

#### *3.1.2.3 Les outils de recherche sélectionnés*

Parmi les outils de recherche permettant de trouver des ressources ontologiques répondant à des besoins de réutilisation figurent les moteurs de recherche sémantique, les portails et les bibliothèques d'ontologies.

- Les moteurs de recherche Swoogle (Finin et al., 2005), Watson (D'Aquin et al., 2011) et Vocab.cc (Stadtmüller et al., 2013) sont dédiés à la recherche d'éléments d'ontologies dont les vocabulaires constituent un sous-ensemble. L'expression de la recherche se fait en langue naturelle avec des mots clés et les résultats produits sont classés. Les modalités de recherche

sont les suivantes : - par URI ; - par mots-clés correspondant à une classe, une relation, un individu, et au périmètre de l'ontologie (Label, Commentaire, etc.) ; - par requêtes sur des triplets[7] ((terme générique, relation, terme) (terme générique, relation, ?), (terme générique relation, ?)) (Sabou et *al*., 2008).

- Le portail Linked Open Vocabularies (LOV) (une version indépendante est hébergée à l'OKFN), un des modules de Datalift (Vandenbussche et Vatant, 2014), qui est destiné à cataloguer, trouver et choisir des ontologies. Chaque vocabulaire est maintenu et publié de manière indépendante tout en entretenant des liens de dépendance vis-à-vis des autres vocabulaires.

- Les bibliothèques d'ontologies (BioPortal, AgroPortal, OBOFoundry, ODP) dont la finalité est de permettre et faciliter l'interopérabilité, fournir des ontologies reconnues et testées. Elles sont caractérisées par leur domaine, leur contenu et leurs fonctionnalités.

- L'outil OntoFox[8] prend en charge la réutilisation des ontologies. Il permet aux utilisateurs de saisir des termes, d'extraire des propriétés sélectionnées, des annotations et certaines classes de termes connexes à partir d'ontologies sources et de sauvegarder les résultats à l'aide de la sérialisation RDF / XML du langage OWL. Ontofox dispose également d'un algorithme d'extraction de termes fondé sur SPARQL qui extrait des termes liés à un ensemble donné de termes de signature. En outre, Ontofox fournit une option pour extraire la hiérarchie enracinée à un terme d'ontologie spécifié.

#### *3.1.2.4 Le modèle proposé pour les outils de recherche*

Le modèle pour les moteurs et le portail LOV a été construit à partir du tableau comparatif établi par (Vandenbussche et al., 2017). Les caractéristiques retenues comportent la méthode d'accès à la ressource (automatique et/ou manuelle), le type de la ressource (Semantic Web Document, concepts, termes de vocabulaire, vocabulaires), la métrique de ranking, le filtrage du domaine, l'accès au service web, etc.).

Le modèle pour les bibliothèques a été établi à partir du tableau comparatif proposé par (Debashis Naskar et Biswanath Dutta, 2016). Les caractéristiques décrites sont le domaine, les fonctionnalités de navigation et de recherche (sujet, structure, langue), le type d'appariement supporté (entre classe ou ontologie), le processus de soumission, l'utilisation et l'accès aux ontologies, les formats de fichiers en entrée et sortie, les outils associés.

### 3.1.3 Outils de validation

#### *3.1.3.1 Les scénarios*

Scénario 1 : L'utilisateur cherche à valider l'ontologie qu'il a construite et s'interroge sur les outils lui permettant de réaliser cette tâche. L'ontologie doit décrire les outils permettant les différents types de validation.

Scénario 2 : Quels sont les critères garantissant la validité d'une ontologie ? oogo doit préciser les critères utilisés pour la validation et les outils leur correspondant.

#### *3.1.3.2 Question de compétences*

Question 1 : Existe-il des outils pour valider les ontologies ? (scénario 1)

Question 2 : Quel outil me permet de vérifier la structure de mon ontologie ? (scénario 1 et 2)

Question 2 : Quel outil me permet de vérifier le modèle de l'expertise représenté dans mon ontologie ? (scénario 2)

[7] http://scarlet.open.ac.uk/poweraqua/index.html

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

#### *3.1.3.3 Les outils de validation sélectionnés*

Nous avons utilisé l'analyse des outils de validation réalisée par (Richard, 2017) pour la sélection des outils de validation. La validation de la qualité d'une ontologie se définit selon deux axes principaux qui sont la validation de la structure et la validation de la sémantique. La validation de la structure correspond à une validation formelle de la logique du modèle. Elle peut être réalisée automatiquement, grâce au développement d'outils dédiés. La validation de la sémantique consiste à mesurer l'adéquation du modèle conceptuel avec la réalité qu'il modélise. Cette étape nécessite une collaboration entre les ontologues et les spécialistes du domaine modélisé dans l'ontologie.

*Les outils de validation de structure*

Les raisonneurs permettent de vérifier la consistance et la cohérence d'un modèle (cf. Les outils de raisonnement, *infra*).

OntoCheck (Schober et *al*., 2012) constitue un module d'extension à l'éditeur d'ontologies Protégé et vise à contrôler le respect des conventions de nommage, ainsi que l'exhaustivité des méta-données. OntoCheck vérifie les cardinalités, la complétude des méta-données, les labels, les conventions typographiques ainsi que les métriques de l'ontologie.

XD Analyzer a été développé dans le cadre du projet NeOn[9], pour faire un retour qualitatif à l'utilisateur en suivant la méthodologie XD. Cette dernière fournit une liste de bonnes pratiques (concernant entre autres, les labels, les commentaires, les concepts non utilisés) à respecter pour la construction d'ontologies. Le module d'extension RaDON (Ji et al., 2009) a été ajouté à Neon Toolkit afin de faciliter les tâches de vérification de la consistance. Il comporte deux plugins permettant de travailler sur une ontologie ou un réseau d'ontologies.

OntOLogy Pitfall Scanner! (OOPS!)[10] (Poveda-Villalón et *al*., 2012) est un outil indépendant de tout éditeur d'ontologies. Il s'utilise uniquement en ligne. Le but de OOPS ! est l'identification des anomalies ou des mauvaises pratiques dans une ontologie. L'application prend en entrée l'URI d'une ontologie ou le code source en RDF. L'ontologie est chargée *via* l'API Jena avant d'être analysée pour en extraire les erreurs potentielles. Le résultat est une page répertoriant les erreurs selon le piège (pitfall) identifié, avec une proposition de résolution de l'erreur. Les pitfalls peuvent concerner des éléments individuels, plusieurs éléments ou toute l'ontologie.

*Les outils de validation sémantique*

Les outils permettant de valider la sémantique d'une ontologie nécessitent le recours à des experts du domaine modélisé par l'ontologie. Des outils supportant le développement de construction collaborative d'ontologies contribue à la validation sémantique.

Le serveur Ontolingua propose un environnement collaboratif pour parcourir, créer, éditer, modifier ou utiliser les ontologies (Farquhar et *al*., 1997)

WebProtégé : a été développé en reprenant l'architecture de Protégé. Il permet également le développement collaboratif d'ontologies et est accessible *via* n'importe quel navigateur web (Tudorache et *al*., 2013).

#### *3.1.3.4 Le modèle pour les outils de validation*

Le modèle des outils de validation décrit les fonctionnalités concernant la validation collaborative d'ontologies et précise les types de validation possible.

### 3.2 Activités de développement

Parmi les activités intervenant dans la phase de développement, nous considérons la conceptualisation, la formalisation, la modularisation, l'alignement. Les outils supportant ces

[9] http://www.neon-project.org

[10] http://oops.linkeddata.es/catalogue.jsp

activités sont les outils d'édition, les raisonneurs, les outils d'alignement et les outils de visualisation.

### 3.2.1 Outils d'édition

Un éditeur d'ontologie est un outil permettant d'inspecter, de parcourir, de codifier, de modifier les ontologies et de soutenir de cette manière le développement de l'ontologie et la tâche de maintenance.

#### *3.2.3.1 Les scénarios*

Scénario 1 : L'ontologue est débutant et cherche un éditeur pour construire une ontologie. Il n'a pas forcément une idée précise de fonctionnalité de l'outil. oogo doit décrire les différentes fonctionnalités de la phase de développement de l'ontologie.

Scénario 2 : L'ontologue a une idée précise de la nature de l'ontologie qu'il souhaite construire et des tâches qu'elle permettra de réaliser (annotation, raisonnement).

#### *3.2.3.2 Les questions de compétence*

Question1 : Quels sont les éditeurs d'ontologie disponibles ? (scénario 1)
Question 2 : Quel est le meilleur éditeur d'ontologie pour un débutant ? (scénario 1)
Question 3 : Quelles sont les activités supportées par l'éditeur d'ontologie ? (scénario 2)

#### *3.2.3.3 Les outils d'édition sélectionnés*

Nous avons utilisé le wiki[11] du W3C pour répertorier les éditeurs les plus utilisés actuellement pour construire des ressources ontologiques au sens où nous les avons définis dans cet article. Par conséquent nous n'avons pas retenu les éditeurs de vocabulaire. Parmi la liste des éditeurs inventoriés nous avons retenu Protégé, NeOn Toolkit, SWOOP et TopBraid Composer.

Nous avons également exploité les comparaisons des différents éditeurs pour identifier les propriétés des outils d'édition et construire un modèle de ces outils (Abburu, 2012 ; Abburu & Babu, 2013 ; Alatrish, 2013 ; Kapoor & Sharma, 2010 ; Parveen et al., 2016 ; Slimani, 2015).

#### *3.2.3.4 Le modèle pour les outils d'édition*

Nous avons retenu les caractéristiques servant à ces comparaisons pour construire le modèle des outils d'édition. Elles concernent l'architecture de l'outil et sa capacité d'évolution, l'interopérabilité, le paradigme de représentation des connaissances et le support méthodologique, les services d'inférences, les services de visualisation, la gestion des versions et l'utilisabilité. Certaines de ces caractéristiques sont déjà décrites dans le modèle du concept Outil (cf. Figure 1).

L'étude de l'utilisabilité des éditeurs réalisée par (Garcia Barriocal et al., 2006) pour des ontologues débutants met en évidence des propriétés relatives à la facilité de prise en main des outils. Certaines propriétés nous semble quantifiables comme accessibilité du langage utilisé à une communauté d'utilisateurs non spécialistes, la visualisation hiérarchique des classes, la navigation d'une classe vers ses instances, le filtrage des classes à partir de critères fixés par l'utilisateur

### 3.2.4 Outils de raisonnement

Une ontologie exprimée en OWL peut être considérée comme un ensemble de formules FOL (First Order logic), appelées axiomes, et par conséquent comme une théorie logique. Le vocabulaire exprimé en OWL est bien défini et sans ambiguïté. Cependant, il y a souvent un décalage entre la représentation visée et la représentation réelle de la connaissance décrite

[11] https://www.w3.org/wiki/Ontology_editors

dans l'ontologie. Alors que la connaissance du domaine d'intérêt est généralement bien maîtrisée ou tout au moins comprise, il n'est pas trivial de prévoir et de comprendre les conséquences logiques d'un ajout, d'un retrait ou d'une modification d'axiome, notamment quand la terminologie est fortement interconnectée. Ainsi, un ontologue a besoin d'utiliser des systèmes de déduction automatisés, communément appelés raisonneurs, pour vérifier certaines formes de déduction à partir d'axiomes préétablis et de rendre explicites ceux qui sont déclarés implicitement. Les raisonneurs sont donc, à l'heure actuelle, des éléments clés pour travailler avec des ontologies (Alaya, 2016).

#### *3.2.4.1 Les scénarios*

Scénario 1 : L'utilisateur cherche l'outil support du raisonneur (Neon, Protégé, Swoop, etc.). L'ontologie doit donc permettre de représenter les différents supports et de leur associer les raisonneurs existants.

Scénario 2 : L'utilisateur cherche à connaître les tâches supportées par un raisonneur. L'ontologie doit décrire les tâches de raisonnement et les entités sur lesquels ils portent.

Scénario 3 : L'utilisateur cherche les capacités d'explication et de gestion des erreurs du raisonneur.

Scénario 4 : L'utilisateur cherche le raisonneur adapté au contenu de l'ontologie à exploiter ? Dépend de la formalisation adoptée pour l'ontologie qui dépend elle-même des tâches à réaliser (annotation, raisonnement). Au différents profils sont associés des raisonneurs

#### *3.2.4.2 Les questions de compétence*

Question 1 : Sous quel système d'exploitation est-il exécutable ? Quel est le type de licence ? (scénario lié au modèle outil)

Question 2 : Avec quel éditeur le raisonneur est-il compatible ? Est-il standalone ?

Question 3 : Le raisonneur peut-il exploiter des règles et quel est le formalisme supporté ?

Question 4 : Le raisonneur prend t-il en charge le raisonnement sur les individus ?

Question 5 : Le raisonneur est-il adapté à la nature de l'ontologie que j'ai construite ?

Question 6 : Quels types de justification (explications sur les inconsistances ) peut fournir le raisonneur ?

Question 7 : Quel type de profil OWL est pris en charge par le raisonneur ?

#### *3.2.4.3 Les outils de raisonnement*

Nous avons utilisé les comparaisons réalisées par (Dentler et *al*., 2011), (Abburu, 2012), (Matenzoglu et *al*., 2016) et (Alaya, 2016) pour construire le modèle des outils de raisonnement.

(Dentler et *al*., 2011) dressent une liste de caractéristiques guidant l'utilisateur dans le choix d'un raisonneur en fonction des besoins d'une application. Elle est envisagée selon deux dimensions : les caractéristiques du raisonnement et l'utilisabilité pratique. La première dimension regroupe : la procédure ou l'algorithme utilisé par le raisonneur pour le raisonnement en logique de description (algorithme Tableau, Hyper-Tableau et Consequence-Based), la robustesse et la complétude (vérifie si toutes les inférences possibles sont réalisées) qui peut aider à une accélération du temps de raisonnement, l'expressivité et la complétude (lien avec les profils OWL), la classification incrémentale (synchronisation du raisonneur), le support de l'utilisation de règles (par exemple, SWRL), la justification (explication pour un concept inconsistant, disjonction de concept, implication logique, etc.), le support des tâches de raisonnement sur la Abox (raisonnement sur les individus, vérification de la consistance de la Abox, etc.). La seconde a trait à la disponibilité ( plugin, standalone), au type de licence et d'autres caractéristiques comprenant le type de code, les plateformes compatibles, l'interface Jena native, etc.

Après avoir décrit l'architecture des raisonneurs raisonneurs Pellet, RACER, FaCT++, Snorocket SWRL-IQ, ELK, HermiT, CEL et TrOWL (Abburu, 2012) en dresse une comparaison selon les caractéristiques définies par (Dentler, 2011).

(Matenzoglu et *al*., 2016) réalise une comparaison de 35 raisonneurs dont la plupart ont été développé entre 2010 et 2015. Elle est fondée sur des critères d'utilisabilité (services de raisonnement supporté, les niveaux d'expressivité et la complétude du raisonneur) et les algorithmes de raisonnement implémentés.

#### *3.2.4.4 Le modèle pour les outils de raisonnement*

Le modèle des outils de raisonnement a été établi à partir du tableau comparatif présenté par (Alaya, 2016) qui est fondé sur l'article (Matenzoglu et al., 2016). Les raisonneurs décrits sont actuellement FaCT++ , Pellet , ELK , HermiT , Konclude , Euler/Eye et NoHR .

### 3.2.5 Outils de conversion de données en OWL

La finalité de ces outils est la conversion de données contenues dans des feuilles de calcul en OWL.

#### *3.2.5.1 Les scénarios*

Scénario 1 : L'utilisateur dispose de documents existant sur la connaissance à formaliser, le plus souvent sous la forme de listes ou de feuilles de calcul. Il souhaite pouvoir les greffer facilement sur le squelette de son ontologie en utilisant les relations qu'il a prédéfinies.

Scénario 2 : L'utilisateur souhaite importer des éléments de hiérarchie à partir d'ontologies existantes avec cependant la possibilité de les amender ou de les complèter.

Scénario 3 : L'utilisateur souhaite mettre à la disposition d'intervenants extérieurs des parties de son travail en cours à des fins de vérification ou de validation.

Scénario 4 : L'utilisateur souhaite introduire dans l'ontologie d'importantes quantités de faits ou de caractéristiques sans avoir la charge de les éditer tous manuellement.

#### *3.2.5.2 Les questions de compétence*

Question 1 : L'outil permet-il l'import de données exprimées dans des tableaux (feuilles de calcul) ?

Question 2 : L'outil permet-il l'opération inverse (export vers des feuilles de calcul) d'extraits sélectionnés de son ontologie ?

Question 3 : Ces fonctionnalités sont-elles disponibles pour une utilisation en mode « ligne de commande » pour supporter un certain degré d'automatisation des tâches répétitives lors de l'enrichissement de son ontologie ?

#### *3.2.5.3 Les outils de conversion de données*

Nous avons sélectionné l'outil Cellfie un plugin Protégé 5 intégrant MappingMaster (O'Connor M. J., 2010) et l'outil FOE (Despres & Guezennec, 2017). Populous et WebPopulous offrent des fonctionnalités similaires ainsi que l'option d'import Excel disponible dans Topbraid Composer .

#### *3.2.5.4 Le modèle pour les outils de conversion de données*

Pour construire le modèle des outils de conversion de données, nous avons le type et d'export et les modalités d'utilisation, le degré d'automatisation autorisé et les constructions qu'il est possible de créer.

### 3.2.6 Outils d'alignement

Pour construire le modèle des outils d'alignement nous avons commencé à tester les outils répertoriés sur le site de AI3[12], la librairie des outils d'alignement hébergée par l'INRIA[13] et le site du projet Ontobeep[14]. Nous inventorions également les outils de visualisation d'alignement tels que (Lanzenberger et *al*., 2011). Cette partie du travail est en cours et ne sera pas présenté dans l'article.

### 3.2.7 Outil de gestion d'évolution

Pour construire le modèle des outils d'évolution nous avons commencé à inventorier les outils existants. Nous travaillons à partir des travaux récents de (Lambrix et *al*., 2016) qui réalisent une synthèse de ces outils du point de vue de la visualisation.

Les outils que nous analysons sont CODEX (Hartung et al., 2012), REX (Christen et *al*., 2015), NeonToolkit (Palma et *al*., 2012).

### 3.2.8 Outils de visualisation

La visualisation d'ontologies est une tâche qui aide à la compréhension des modèles représentés. Elle est utile à l'ensemble des acteurs participant à sa construction. Or, il n'est pas simple de créer une visualisation des ontologies en raison de la complexité de ces ressources. Les entités à visualiser sont la hiérarchie des concepts, les relations entre les concepts, les attributs des concepts et les individus relevant des concepts représentés (Katifori, 2007). Nous avons identifié deux publications faisant références à des ontologies VO formalisant les connaissances utiles la visualisation d'ontologies (Shu, 2007) et (M. Voigt & J. Polowinski, 2011) qui décrivent les primitives utilisées dans le domaine de la visualisation. Elles ne sont pas directement exploitables mais nous avons réutilisé certaines des classes modélisées.

#### *3.2.8.1 Les scénarios*

Scénario 1 : L'utilisateur souhaite visualiser la hiérarchie des concepts.

Scénario 2 : L'utilisateur souhaite se centrer sur un concept afin de visualiser globalement l'ensemble des liens qu'il entretient avec le réseau de concepts.

Scénario 3 : L'utilisateur souhaite visualiser les explications fournies par le raisonner et disposer d'une représentation graphique.

#### *3.2.8.2 Les questions de compétence*

Question 1 : Quel outil permet de visualiser l'ontologie ?

Question 2 : Quelles sont les entités visualisable de l'ontologie ?

Question 3 : Quelles sont les actions possibles pour visualiser l'ontologie ?

#### *3.2.8.3 Les outils de visualisation*

Les outils décrits dans la version actuelle de oogo sont GraphViz, CropCircles, KC-Viz, Knoocks, LogDiffViz, OntoGraph, OntologyVisualizer, Jambalaya, OntoTGVizTab, VOWL, WebVOWL, Protupos.

#### *3.2.8.4 Le modèle pour les outils de visualisation*

Le modèle des outils de visualisation a été établi à partir de la comparaison et l'évaluation des visualisations d'ontologies présentées par (Balzer et al., 2015) et des publications relatives à certains outils (Kuhar & Podgorelec, 2012), (Lohmann et al., 2016).

[12] http://www.mkbergman.com/1769/50-ontology-mapping-and-alignment-tools/

[13] http://alignapi.gforge.inria.fr/impl.html

[14] http://www.ontobee.org/tutorial/ontobeep#features

Dans le cadre de la construction du service web de visualisation centré utilisateur Protupos, (Despres &Nobecourt, 2018) ont caractérisé les outils de visualisation actuellement utilisés :

- Graphviz est un plugin de Protégé qui s'exécute dans son propre onglet. Il sert principalement à la visualisation graphique d'ontologie grâce à l'utilisation d'un graphe de nœuds et d'arcs qu'il est possible d'étendre et de réduire dynamiquement via des actions sur des boutons. Sur le principe, Graphviz est très utile pour parcourir la hiérarchie et la manipuler (expansion, respectivement réduction, des propriétés relatives à un nœud). Il permet également la visualisation de la hiérarchie des classes inférées.
- *VOWL* est un plugin de Protégé qui s'exécute dans son propre onglet et qui ne prend pas en compte les actions effectuées par ailleurs. Par exemple, il n'est pas possible de se centrer sur un concept et de naviguer autour de ce dernier dans l'onglet VOWL. Sur le principe, VOWL est une spécification de notations graphiques pour les ontologies écrites en OWL http://purl.org/vowl/spec/. L'outil est clairement tourné vers l'ontologue.
- WebOWL (http://visualdataweb.de/webvowl/) est un service web implémentant VOWL indépendamment de Protégé. Les actions peuvent être effectuées grâce aux boutons et aux menus. Un inconvénient majeur de WebOWL est que comme pour tout service web, l'ontologie doit être téléchargée.
- CropCircles est un outil intégré à SWOOP qui permet de visualiser la hiérarchie des concepts. Un concept est représenté par un cercle, les fils de ce concept sont représentés à l'intérieur de ce cercle et la hiérarchie est visualisé selon un représentation en coupe. Il est possible de zoomer sur une couche particulière de la hiérarchie en cliquant dans le cercle et de naviguer dans la hiérarchie en cliquant à l'extérieur du cercle. La vision en coupe par niveau hiérarchique permet notamment de bien appréhender la densité d'un sous-arbre par rapport à un autre.

Le développement de Protupos a permis de définir des fonctionnalités que nous avons utilisées pour construire le modèle des outils de visualisation :

- visualisation interactive de la hiérarchie : utilisation d'une représentation circulaire zoomable par actions de l'utilisateur ;
- de la hiérarchie par coupe de niveau ;
- visualisation de chaîne de propriétés centrée sur un nœud avec la possibilité de développer dynamiquement la construction de la chaîne ;
- visualisation sous la forme de réseaux d'un type de propriété ;
- visualisation centrée sur un concept en utilisant une fonction « élastique (permettant de modifier le focus) » ;
- nuage de tags associés à un concept.

## 4. Présentation de oogo et évaluation

oogo a été éditée avec le logiciel Protégé (version 5.2.0)[15]. L'ontologie et la description de ses ressources et propriétés seront publiées selon les principes des données liées sur le web et le schéma sera identifié par l'URI HTTP https://www-limics.smbh.univ-paris13.fr/oogo#. Plusieurs vocabulaires sont utilisés pour la construction du modèle outil : dcterms, doap, dvia, foaf, org. Dans cette version de oogo, nous ne pouvons pas réutiliser dvia:url car le type anyUri n'est pas directement supporté par Protégé 5.

[15] https://protege.stanford.edu/

## 4.1 Métrique de oogo

Metrics

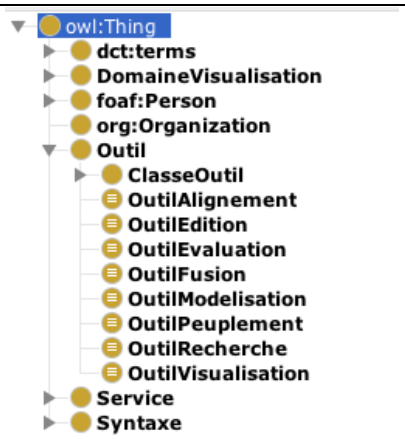
| | |
|---|---|
| Axiom | 756 |
| Logical axiom count | 368 |
| Declaration axioms count | 208 |
| Class count | 155 |
| Object property count | 17 |
| Data property count | 12 |
| Individual count | 23 |
| Annotation Property count | 6 |
| DL expressivity | ALCHO(D) |

FIGURE 3 – Métrique de l'ontologie oogo

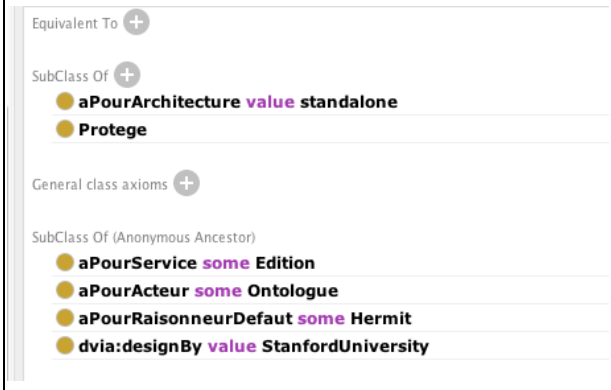
La métrique de l'ontologie est présentée Figure 3. Elle comporte actuellement 155 classes, 17 object property et 12 data property. Elle est en cours d'évolution.

## 4.2 Aperçu des classes de oogo

Nous présentons une description du concept Outil dans oogo. Dans la classe ClasseOutil l'ensemble des outils étudiés sont décrits comme des sous-classes. Par exemple, la classe Protégé comporte les sous-classes Proétég3-x, Protégé4-x, Protégé5-x. Des individus relèvent de ces différentes sous-classes. Elle est également définie comme disposant d'un service d'édition. Les classes définies, comme OutilEdition de la figure 4(A), permettent d'utiliser le raisonneur pour caractériser les outils selon les services qui leur correspondent.




(A)




(B)

Figure 4 – Vue globale de oogo (A) et Définition d'un outil (B)

### 4.3 Requêtes DL Query

Nous montrons par le biais de quelques exemples de requêtes DL Query que les questions de compétence trouvent une réponse.

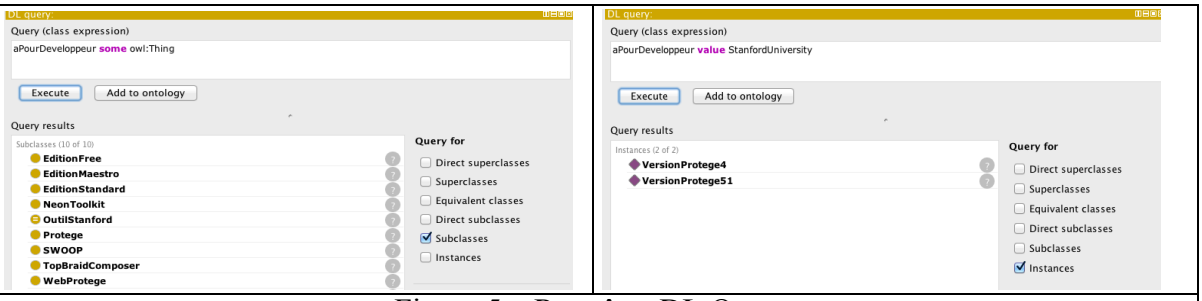



Figure 5 – Requêtes DL Query

## Conclusion

Cet article présente un travail en cours qui répondait à une demande qui nous a permis de prendre conscience de la nécessité de centraliser l'expérience de la communauté de manière synthétique afin de répondre simplement et rapidement à des questions pratiques concernant la disponibilité des différents outils utilisables dans le cadre de la construction d'ontologies. L'ontologie ébauchée dans cet article est par nature destinée à évoluer régulièrement une fois qu'elle aura été mise en ligne dans sa version initiale. Actuellement, la langue utilsée pour le développement de oogo est le français néanmoins l'ensemble des entités sont décrites avec des labels et altlabel en anglais. Nous avons le projet de rendre sa construction collaborative lorsqu'elle sera publiée afin de s'assurer qu'elle ne devienne pas obsolète. De cette façon, les retours d'expérience pourront être valorisés et bénéficier à l'ensemble de la communauté.

## Références


ABBURU, S. (2012). A Survey on Ontology Reasoners and Comparison. International Journal of Computer Applications, 57(17):33–39.

ABBURU, S., BABU, G. S. (2013). Survey on Ontology Construction Tools. International Journal of Scientific & Engineering Research, Volume 4, Issue 6, pp.1748-1752.

ALATRISH E. S. (2013). Comparison Some of Ontology Editors, in Management Information Systems, vol 8, n°2, pp. 018-024, 2013.

ALAYA N. (2016). Managing The Empirical Hardness of The Ontology Reasoning using The Predictive Modelling. Thèse de l'université de Tunis El Manar et de l'université Paris8.

BALZER L., DO M.T., MASELUK D. (2015). Comparison and Evaluations of Ontology Visualizations. Rapport de recherche, H.5.2, I.2.4.

CHRISTEN V., GROSS V., HARTUNG M. Region Evolution eXplorer a tool for discovering evolution trends in ontology regions. Journal of Biomedical Semantics, 6:Article 26, 2015.

D'AQUIN M., MOTTA E. (2011). Watson, more than a Semantic Web search engine. Semantic Web 2(1): 55-63.

DENTLER K., CORNET R., & ten TEIJE A., & de KEIZER N. (2011). Comparison of reasoners for large ontologies in the OWL 2 EL profile. Semant. web, 2(2):71–87.

DESPRES S., GUEZENNEC G. - Flat OWL Editor : un outil utilisant des feuilles de calcul pour découpler les tâches des acteurs impliqués dans la gestion d'une ontologie. Toth2017, (à paraître).

DESPRES S., NOBECOURT J.- Quelles fonctionnalités pour un outil de visualisation d'ontologie ? Atelier VIF Visualisation d'informations, Interaction, et Fouille de données EGC 2018 (http://gt-vif.polytech.univ- nantes.fr/egc-vif2018/atelier_VIF_EGC2018.pdf

FININ T., DING L., PA R., JOSHI A., KOLARI P. (2005). Java A., Peng Y. Swoogle: Searching for knowledge on the semantic web. In Anthony Cohn, editor, Proceedings of the 20th National Conference on Artificial Intelligence - Volume 4, AAAI'05, pages 1682–1683. AAAI Press.

GARCIA-BARRIOCANAL E., SICILIA M. A. & SANCHEZ-ALONSO S. (2006). Usability Evaluation of Ontology Editors. Knowledge Organization 32(1),1-9.

HARTUNG M., GROSS A., RAHM E. CODEX: Exploration of semantic changes between ontology versions. Bioinformatics, 26(6):895–896, 2012.

HEON M., NKAMBOU R., LANGHEIT C. (2016). Toward G-OWL: A graphical, polymorphic and typed syntax for building formal OWL2 ontologies. Proceedings of the 25th International Conference Companion on World Wide Web. International World Wide Web Conferences Steering Committee.

JAIN V., PRASAD S.V.A.V. (2016). Evaluation and Validation of ontology using Protégé Tool. In International Journal of Research in Engineering & Technology (IMPACT: IJRET) ISSN (E): 2321-8843; ISSN (P): 2347-4599 Vol. 4, Issue 5, May 2016, 1-12.

KAPOOR, B., SHARMA, S.A. (2010). Comparative Study Ontology Building Tools for Semantic Web Applications. International journal of Web & Semantic Technology (IJWesT), 1(3).

KATIFORI A., HALATSIS C. LEPOURAS G., VASSILAKIS C., GIANNOPOULOU E. (2007). Ontology Visualization Methods – A Survey. ACM Computing Surveys, Vol. 39, N°4, Article 10.

LAMBRIX P., DRAGISIC Z., IVANOVA V., ANSLOW C. Visualization for Ontology Evolution, VOILA 2016 Visualization and Interaction for Ontologies and Linked Data, 2016, pp.54-67.

LANZENBERGER M., SAMPSON J. (2011). Ontology Alignment Quality: A Framework and Tool for Validation. IJISMD 2(3): 1-23

LOHMANN S., LINK V., MARBACH E. & NEGRU S. (2016). Visualizing ontologies with VOWL.

7(4),21. Already accepted papers for the EKAW 2014 special issue, extended version.

MATENTZOGLU N., Leo J., Hudhra V., SATTLER U., & Parsia B.. (2015). A survey of current, stand-alone OWL reasoners. In Proceedings of the 4th International Workshop on OWL Reasoner Evaluation, pages 68–79.

NASKAR D., DUTTA B. (2016) - Ontology Libraries: A Study from an Ontofier and an Ontologist Perspectives. 19th International Symposium on Electronic Theses and Dissertations, Lille.

O'CONNOR M. J., HALASCHEK-WIENER C., MUSEN M. A. (2010). Mapping master: a flexible approach for mapping spreadsheets to OWL. In ISWC'10 Proceed-ings of the 9th international semantic web conference on The semantic web - Volume Part II, 194-208.

PALMA R., ZABLITH F., HAASE P., CORCHO O. Ontology evolution. In MC Suarez-Figuero, A Gomez-Prez, E Motta, and A Gangemi, editors, Ontology Engineering in a Networked World, pages 235–255. 2012.

PARVEEN, KUMAR SAHNI D.K., KHURANA D., NANDAL R. (2016). Ontology Development Tools and Languages: A Review. Vol. 5, Issue 6, June, 2016

POVEDA-VILLALON M., M.C., GOMEZ-PEREZ A, POVEDA-VILLALON M. (2014). OOPS! (OntOlogy Pitfall Scanner!): An On-line Tool for Ontology Evaluation. Int. J. Semantic Web Inf. Syst. 10(2): 7-34.

RICHARD M. (2017). Apports de la modélisation ontologique pour le partage des connaissances en psychiatrie. Thèse de l'université Pierre et Marie Curie - Paris VI.

SABOU M., D'AQUIN M., MOTTA E. (2008). SCARLET: SemantiC RelAtion DiscoveRy by Harvesting OnLinE OnTologies. ESWC 2008: 854-858.

SEONGWOOK Y., ANCHIT A., PREETHAM C., PAAVANY J., ASHISH M. and SHIKHA S. (2009). Survey about Ontology Development Tools for Ontology-based Knowledge Management, University of Southern California.

SLIMANI, T. (2015). Ontology Development: A Comparing Study on Tools, Languages and Formalisms. In Indian Journal of Science and Technology, Vol 8(24), doi:10.17485/ijst/2015/v8i34/54249.

STADTMÜLLER S., HARTH A., GROBELNIK M. (2013). Accessing information about linked data vocabularies with vocab.cc. In Juanzi Li, Guilin Qi, Dongyan Zhao, Wolfgang Nejdl, and Hai-Tao Zheng, editors, Semantic Web and Web Science, Springer Proceedings in Complexity, pages 391–396. Springer New York. doi:10.1007/978-1-4614-6880-6,34.

SHU G., AVIS N.J. (2008). Bringing semantic to vizualization services. Advances in Engineering Software 39. Pp.514-520.

STEIGMILLER A., LIEBIG T., & BIRTE. G. (2014).Konclude: System description. Web Semantics: Science, Services and Agents on the World Wide Web, 27(1).

STOJANOVIC L., MOTIK B. (2002) Ontology Evolution within Ontology Editors. In Conference on the Evaluation of Ontology-based Tools.

STUDER R., BENJAMIN V. R et FENSEL. (1998). D. Knowledge Engineering: Principles and methods. Data & Knowledge Engineering 25, pp. 161-197.

SUAREZ-FIGUEROA M.C., GOMEZ-PEREZ A., FERNANDEZ-LOPEZ M. (2015). The NeOn Methodology framework: A scenario-based methodology for ontology development. Applied Ontology 10(2): 107-145.

SUAREZ-FIGUEROA M.C., GOMEZ-PEREZ A. (2008). Towards a Glossary of Activities in the Ontology Engineering Field, LREC.

TSARKOV D. & HORROCKS I. (2006). Fact++ description logic reasoner: System description. In Proceedings of the Third International Joint Conference on Automated Reasoning, pages 292–297.

TUDORACHE T., NYULAS C., NOY, N. F. & MUSEN, M. A. (2013). Webprotégé : A collaborative ontology editor and knowledge acquisition tool for the web. Semantic web, 4(1) :89–99. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3691821/.

USCHOLD M., GRUNINGER M. (1996). Ontologies: Principles, methods and applications. The knowledge engineering review, vol. 11, no 2, p. 93–136.

VANDENBUSSCHE P.Y. & VATANT B. (2014). Linked Open Vocabularies. ERCIM news, 96:21–22.

VANDENBUSSCHE P.Y., ATEMEZING G., POVEDA-VILLALON M., VATANT B. (2017). Linked Open Vocabularies (LOV): A gateway to reusable semantic vocabularies on the Web. Semantic Web 8(3): 437-452.

VOIGT M. & POLOWINSKI J. (2011). Towards a Unifying Visualization Ontology. Technical Report – Technische Universität Dresden (3/2011).

# Représentation des Connaissances dans un Langage Visuel Typé : une Première Approche

Florence Dupin de Saint-Cyr[1], Denis Parade[2]

[1] IRIT, Université de Toulouse, Toulouse, France
Florence.Bannay@irit.fr

[2] SCÉNARIO INTERACTIF, Toulouse, France
infos@scenario-interactif.com

**Résumé** : Cet article présente un ensemble de principes qu'un langage visuel de représentation des connaissances doit satisfaire. Nous avons introduit quatre principes formel (accessibilité, navigation, efficacité d'accès, existence d'un point d'entrée) et sept postulats en langage naturel visant à contraindre la cohérence, compacité, navigabilité du langage par rapport aux attributs visuels (position, forme et couleur).

Après une présentation du langage MOT, nous introduisons le langage VTL (Visual Typed Language) qui en est inspiré. VTL est un langage visuel typé qui satisfait la plupart des principes que nous proposons.

Nous avons séparé les éléments du langage VTL en trois catégories : *entités*, *actions* et *conditions*. Les expressions de chaque catégorie sont soit des éléments génériques soit des instances. Ils sont reliés entre eux par des attributs dépendant de leur catégorie, ces liens permettent la navigation.

Enfin, nous avons réalisé une maquette interactive à l'aide du logiciel XMind, initialement dédié aux cartes mentales.

**Mots-clés** : langage visuel, perception visuelle humaine, interactivité, navigation par zoom avant/arrière

## 1 Introduction

De nos jours, la transmission des idées passe principalement par la langue écrite. Comme tout langage, l'écriture doit répondre à une syntaxe et à des règles grammaticales précises. Par convention, et quelle que soit la langue utilisée, des mots composent des phrases et sont posés séquentiellement sur des lignes qui vont également guider l'œil lors de la lecture. Même si cela nous permet d'exprimer des idées complexes et subtiles, certaines représentations vont au-delà du mode linéaire et ajoutent une distribution spatiale à l'information, éventuellement sur un support interactif, ou via une approche combinant une vue globale et une vue détaillée.

Pour aller au delà de l'utilisation du texte, il y a près d'un siècle, Otto Neurath (Neurath, 1936) a entrepris des recherches pour définir un langage graphique universel. Ce qui l'a conduit à Isotype (Système International de Typographie Picture Education). La sémiotique des pictogrammes était théorisée afin de définir une communication visuelle par des symboles graphiques compris par tous. Cette idée d'utiliser des images pour représenter et transmettre des informations a aussi été largement exploitée dans différents domaines. En effet, les pictogrammes sont des vecteurs d'information très efficaces qui sont utilisés dans les panneaux de signalisation routière par exemple. Selon Teboul (Teboul & Damier, 2017), les êtres humains ont la capacité de traduire immédiatement une forme graphique en un élément sémantique. De plus, la perception visuelle humaine est bien adaptée pour appréhender une situation, reconnaître un lieu et par extension une représentation graphique dans son ensemble. Notre cerveau est en effet particulièrement efficace pour traiter rapidement une information visuelle (on admet [1] souvent que 90% des informations qui arrivent au cerveau sont sous cette forme). Aujourd'hui, il a été montré que les informations présentes sur Internet et les réseaux sociaux

1. Malheureusement, autant que nous le sachions, cette affirmation est toujours utilisée sans aucune référence à une étude scientifique sur ce sujet...

ont beaucoup plus de chances d'être comprises, mémorisées et partagées si elles comportent des éléments visuels (animés ou non).

Il existe déjà beaucoup de représentations visuelles utilisées dans la vie de tous les jours : les cartes mentales introduites par Tony Buzan (Buzan, 1974), les cartes topiques (Pepper, 2000), les cartes conceptuelles (Novak & Cañas, 2008) les diagrammes de Venn (Novak & Cañas, 2008), les frises chronologiques, les organigrammes de programmation (ISO, 1985), les cartes géographiques, etc. Elles ont toutes des aspects intéressants : les diagrammes de Venn sont faciles à comprendre. Les frises chronologiques et les cartes géographiques ont un fort ancrage culturel donc sont également compris immédiatement. Les organigrammes de programmation apportent une dimension ordonnée (temporelle et/ou hiérarchique) et ne sont pas une simple représentation statique. Les cartes mentales offrent une grande liberté de création, sont faciles à mettre en œuvre, et facilitent la mémorisation. Quant aux cartes conceptuelles, elles permettent à l'utilisateur de décrire un mécanisme ou une procédure sans ambiguïté. Cependant, elles ont toutes quelques inconvénients : pour les cartes mentales et les cartes topiques, il peut y avoir des ambiguïtés dans l'utilisation des mots-clés et des relations que l'utilisateur est libre de définir sans aucune contrainte. Les cartes conceptuelles sont graphiquement assez pauvres, et le sujet principal est rarement mis en évidence. Une carte conceptuelle mal composée devient rapidement illisible. Les diagrammes de Venn sont d'une utilisation très limitée, peut-être parce qu'ils sont trop simples. Les frises chronologiques nécessitent d'avoir un support physique de la bonne longueur. Il n'y a aucune possibilité de zoomer ou de modifier l'échelle linéaire. Les organigrammes exigent la connaissance des formes de base et n'utilisent pas de couleurs ou de dessins. Souvent, les projections cartographiques déforment la réalité.

Il existe également beaucoup de langages de représentation visuelle moins utilisés du grand public car destinés à des applications très spécifiques, la plupart du temps informatiques. C'est le cas par exemples des « graphes conceptuels » (Sowa, 1984) destinés au codage visuel de phrases et de discours dont on veut capturer la logique, des méthodes SADT (aussi appelée IDF0) (Dickover *et al.*, 1977; Marca & McGowan, 1987), ODT (Rumbaugh *et al.*, 1991) puis UML (D'Souza & Wills, 1998; Rumbaugh *et al.*, 2017) pour la conception de systèmes d'information et plus généralement de tous les formalismes de réseaux sémantiques (voir (Sowa, 2006) pour un panorama). Notons aussi l'existence de langages permettant de représenter des hiérarchies comme OWL (Lamy *et al.*, 2013), ou à la cartographie des connaissances des organisations (avec entre autre le besoin de localiser les connaissances critiques) comme GAMETH (Grundstein, 2002; Grundstein & Rosenthal-Sabroux, 2004). Aucun de ces langages n'a pour objectif principal d'être facilement appréhendé par un lecteur humain, ils demandent donc tous un apprentissage plus ou moins long.

Afin de surmonter les inconvénients des représentations visuelles existantes, nous avons commencé une démarche de formalisation de principes permettant de caractériser une représentation visuelle intuitive et efficace. Nous pensons qu'en imposant des principes rationnels basés sur des fondements en psychologie cognitive, nous pourrons certifier ou non qu'un langage visuel est adapté à la perception et compréhension humaine. Cette première étape nous a amené à définir formellement quatre principes et à en exprimer huit autres sous forme de postulats en langage naturel. De plus, nous proposons un nouveau langage appelé VTL (Visual Typed Language), langage de représentation typé visuel, qui a été créé dans l'optique de satisfaire ces postulats. Ainsi, VTL combine l'utilisation de mots-clés avec des icônes, des images, des schémas et des liens, qui aideront à saisir rapidement le sens de ce qui est exprimé. Cette combinaison d'éléments est liée à la théorie du double codage (Paivio, 1990) : le codage d'un stimulus de deux façons différentes augmente la chance de s'en rappeler par rapport au codage de ce stimulus d'une seule façon. Notre idée est d'enrichir le modèle des cartes mentales afin d'obtenir un schéma qui admet une traduction automatique non ambiguë, qui se base sur l'aspect visuel afin de représenter les propriétés des objets et leurs liens. Notre objectif est que l'existence de cette traduction unique puisse constituer un moyen de comprendre, de raisonner et de décider visuellement.

Une version anglaise de cet article (moins développée par rapport à l'état de l'art) est publiée dans (Dupin De Saint Cyr & Parade, 2018).

## 2 Principes associés à un langage de représentation visuel

Nous allons considérer qu'un langage visuel définit un ensemble d'expressions possibles, ces expressions doivent contenir des informations élémentaires reliées selon le formalisme du langage visuel. Plus précisément, nous considérons un ensemble $S$ de symboles visuels, un ensemble $W$ de mots anglais (les éléments de l'ensemble $S \cup W$ sont appelés « items »), et un ensemble $C$ de connecteurs (avec leur arité). Une expression $e$ d'un langage visuel $L$ est une combinaison d'items (éléments de $S$ et de $W$) à l'aide d'un ensemble $C$ de connecteurs. Les items de $e$ sont notés $items(e)$. La particularité d'un langage visuel est que toute expression $e$ est associée à des attributs visuels tels que forme/couleur/position qui sont attachés à chacun de ses items et également à chacun de ses connecteurs. Nous notons $c(i,j) \in e$ le fait que les deux items $i$ et $j$ sont reliés par le connecteur $c$ dans l'expression $e$. L'existence d'un chemin [2] de longueur $k$ de $i$ à $j$ dans $e$ est notée $path_k(i,j) \in e$. La distance entre deux items dans $e$ est la longueur du plus court chemin entre ces deux items : $dist_e(i,j) = \min\{k | path_k(i,j) \in e\}$

De plus, toute expression $e$ d'un langage visuel a au moins un item qui est considéré visible d'emblée, cet item s'appelle une entrée de $e$, l'ensemble des entrées de $e$, est noté $entrySet(e)$.

Afin de formaliser l'efficacité d'accès aux items d'une expression, nous définissons la mesure suivante :

**Définition 1**
*L'*inefficacité d'accès dans une expression $e$ *est le pourcentage obtenu en divisant la plus grande distance à parcourir pour atteindre un item de* $e$ *depuis une entrée quelconque de* $e$*, par le nombre total d'items dans* $e$*.* $\forall e \in \mathcal{L}$*,*

$$\mathcal{I}(e) = \frac{\max_k\{x \in entrySet(e), y \in items(e), dist_e(x,y) = k\}}{\mid items(e) \mid}$$

*La* mesure d'inefficacité d'accès aux items *d'un langage de représentation visuel* $\mathcal{L}$ *est l'inefficacité maximum qui peut exister dans une de ses expressions*

$$\mathcal{I}(\mathcal{L}) = \max_{e \in \mathcal{L}} \mathcal{I}(e)$$

**Exemple 1**
*Un exemple de représentation visuelle 100% inefficace pour l'accès à l'information, est un texte écrit dans lequel les mots sont les items et leur connexion est donnée par leur ordre dans le texte. Si nous considérons que l'entrée est le premier mot alors la plus grande distance a la taille du texte dans le pire cas où nous voulons accéder au dernier mot du texte depuis l'entrée.*

*Ainsi si le langage contient 35 000 mots (comme le français par exemple) alors l'expression la plus inefficace sera une expression dans laquelle on a écrit les 35 000 mots consécutivement (éventuellement dans un ordre aléatoire), la distance maximale sera donc de 34 999 dans le cas où l'entrée est le premier mot (c'est la distance entre le premier et le dernier mot).*

Dans cette section, nous proposons d'énumérer un ensemble de postulats qu'un langage visuel efficace et bien adapté au fonctionnement humain devrait satisfaire. Nous donnons d'abord 4 postulats qui s'expriment facilement dans les notations que nous avons présentées.

— ***Accessibilité*** :$\forall e \in \mathcal{L}, \forall i \in items(e), \exists k \geq 0, \exists x \in entrySet(e)$ t.q. $path_k(x,i) \in e$.
— ***Navigation*** : $\forall e \in \mathcal{L}, \forall i,j \in items(e)$ si $\exists k \geq 0, \exists x \in entrySet(e)$ avec $path_k(x,i) \in e$ alors $\exists k' \geq 0$ t.q. $path_{k'}(i,j) \in e$.
— ***Efficacité d'accès*** : $\mathcal{I}(\mathcal{L}) < 100\%$.

2. Nous utilisons la définition classique d'un chemin dans un graphe, ici les arcs sont les connexions et les sommets sont les items, avec la convention qu'un chemin de longueur nulle (appelé chemin vide) existe de n'importe quel item à lui-même.

— ***Entrée*** : $\forall e \in \mathcal{L}, \quad |entrySet(e)| = 1$.

En d'autres termes, l'*Accessibilité* impose que toute information exprimée devrait être accessible à l'utilisateur. *Navigation* exprime le fait que depuis n'importe quelle information accessible, l'utilisateur peut accéder à n'importe quelle information exprimée. L'*Efficacité d'accès* impose que n'importe quelle information doit être facilement accessible, c.-à-d., on ne doit pas avoir à lire tout le document afin de la trouver. *Entrée* exige que toute expression a un point d'entrée unique.

Les propriétés suivantes sont des conséquences de ces 4 principes :

**Proposition 1**
*Si un langage $\mathcal{L}$ satisfait Accessibilité et Entrée alors le graphe des connexions de toute expression $e \in \mathcal{L}$ est connexe.*

*Navigation* et *Accessibilité* impliquent que de n'importe quelle position il doit toujours être possible de revenir à une position déjà examinée précédemment :

**Proposition 2**
*Si un langage $\mathcal{L}$ satisfait Navigation et Accessibilité alors la relation de connexion est symétrique.*

Notons que la distance de n'importe quel item au point d'entrée (quand l'entrée est unique) joue un rôle important par rapport à l'*Efficacité d'accès*.

**Proposition 3**
*Soit $\mathcal{L}$ un langage satisfaisant Entrée et Accessibilité et Navigation, $\forall e \in \mathcal{L}$, notons $x_e$ l'unique élément de $entrySet(e)$, si $x_e$ est le centre de gravité de $items(e)$ (par rapport à la distance de $x_e$ aux items) alors $\mathcal{I}(e) < 100\%$.*

Les postulats suivants sont écrits en langage naturel puisqu'ils nécessiteraient d'avantage de notations pour les notions impliquées, l'écriture formelle de ces postulats sera l'objet d'une étude future plus poussée.

— ***Correspondances de similarités*** : Les similarités des positions/formes/couleurs doivent avoir des correspondances en termes de proximité de certains attributs des éléments représentés.
— ***Significativité*** : les positions/formes/couleurs des items ont une signification claire (le centre est important, être à gauche et à droite peut se rapporter à certaine contrainte de priorité, etc.)
— ***3D*** : le langage doit employer les 3 dimensions (environnement naturel des êtres humains), par conséquent la représentation en 2 dimensions devrait être combinée avec une possibilité de zoom avant/arrière.
— ***Brièveté, puissance suggestive et clarté des symboles*** : chaque symbole doit être court et simple. « Court » signifie moins de sept éléments, selon les propriétés de la mémoire temporaire de travail (Miller, 1956). La puissance suggestive pourrait être mesurée en employant la théorie du « double codage » (dual-coding theory (Paivio, 1969)).
— ***Limitation de la surcharge cognitive*** : le nombre d'éléments présentés simultanément à l'utilisateur doit être limité (selon la théorie de Sweller (Sweller, 1994)).
— ***Facile à écrire*** : des outils conviviaux doivent être disponibles,
— ***Traduisible*** : toute expression visuelle valide doit être traduisible dans un formalisme logique et réciproquement.
— ***Cohérence vérifiable*** : il doit y avoir des règles pour vérifier si n'importe quelle expression a au moins une traduction valide.

Notez que tous ces axiomes sont liés à d'autres principes importants et souhaitables. En effet, nous pourrions définir un principe de ***Voisinage*** disant que : Toutes les informations ayant des propriétés communes doivent être proches en distance. Ce postulat implique le postulat *Correspondance de similarités* relativement à la position.

Le postulat *Brièveté, puissance suggestive et clarté des symboles* implique que le langage est ***Facile à lire*** il signifie que les symboles sont compréhensibles sans apprentissage

préalable et que les expressions seront faciles à comprendre et mémoriser (Paivio, 1969). Le postulat *Significativité* peut impliquer que le centre de la représentation a une importance. *Traduisible* implique qu'une expression a une signification univoque et permet d'employer des mécanismes d'inférence.

Notre but est d'établir un langage visuel qui satisfait le plus grand nombre de ces postulats. Nous commençons par rappeler la définition du langage visuel MOT puis nous introduisons le nouveau langage VTL, nous terminons en montrant les principes qu'il satisfait.

## 3 Le langage visuel « MOT »

### 3.1 Description

Nous avons basé la construction de notre langage VTL sur l'approche MOT « Modélisation par objets typés » proposée par Paquette (Paquette, 2010). Cette méthode de représentation des connaissances est dédiée à des formateurs qui doivent définir des systèmes d'apprentissage ou des systèmes d'aide à la réalisation de tâches. Cette approche est basée sur un formalisme graphique.

Selon son auteur, les buts de MOT sont :

— simplicité d'usage pour des personnes non aguerries aux techniques de modélisation des connaissances,
— représentations adaptées à une grande variété de situations et de domaines,
— vue claire des relations entre connaissances, permettant de couvrir tout un domaine sémantique.

Les deux principes fondamentaux de MOT sont :

1. Toute connaissance peut être représentée par un élément dont la forme dépend d'un des trois types de connaissance : déclarative, action ou stratégique. La bordure de l'élément est soit en trait plein pour un élément abstrait soit en traits discontinus pour un élément factuel :

| Type de connaissance | Abstraite | Factuelle |
|---|---|---|
| Déclarative | Concept | Exemple |
| Action | Procédure | Trace |
| Stratégique | Principe | Énoncé |

2. Il y a 6 types de liens entre connaissances : instanciation (I), spécialisation (S), composition (C), précédence (P), Intrant/Produit (I/P) et régulation (R). Tout lien qui ne relève pas de ces 6 types peut être représenté en utilisant un attribut interne.

Voici une description plus précise des 6 types de liens dans MOT :

— I : toute connaissance abstraite (concept, procédure, principe) peut être liée à une de ses instances (connaissance factuelle)
— S : tout concept/procédure/principe abstrait peut être organisé en hiérachies
— C : tout attribut s'il est assez complexe peut être externalisé en un nouveau schéma relié au premier par un lien de composition
— P : Toute procédure peut être décomposé en sous-procédures reliées entre elles par des liens de précédence
— I/P : La notion de procédure peut admettre des entrées/sorties (Intrant/Produit) qui sont des instances ou des concepts abstraits selon le niveau de généralité de la procédure
— R : Des connaissances peuvent régir ou contrôler d'autres connaissances grâce au lien de régulation.

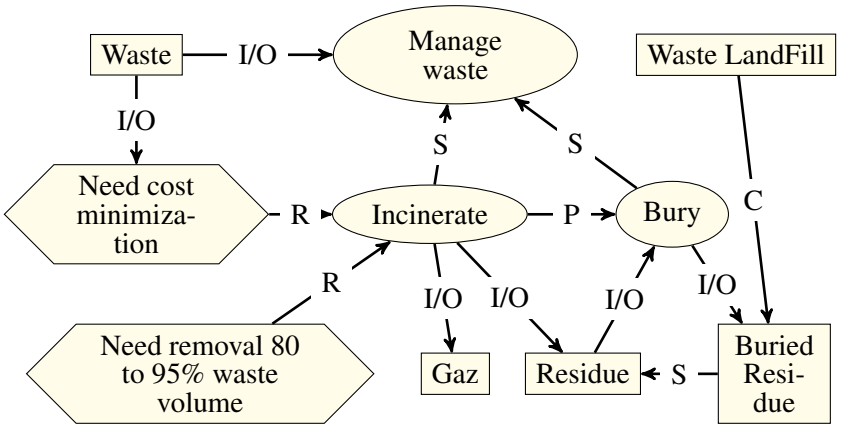



FIGURE *1 – « Manage Waste » dans le formalisme MOT*

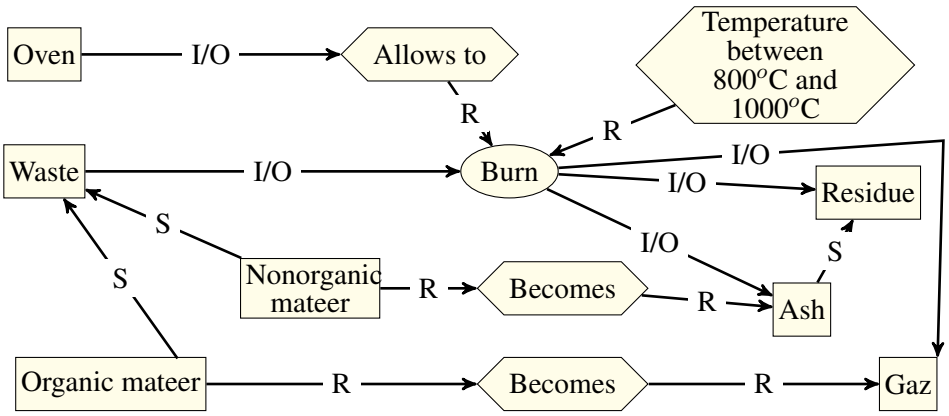



FIGURE *2 – Sous-modèle « Incinerate » dans le formalisme MOT*

Les exemples présentés sur les figures 1 et 2 sont issus de wikipedia (Wikipédia, 2017), ils utilisent MOT pour représenter la procédure « Gérer les déchets », (cet exemple est présenté en anglais afin de mieux pouvoir le comparer avec son équivalent en VTL réalisé dans cette langue, notons que les liens I/P sont notés I/O (pour input/output)).

Dans le travail original de Paquette, il impose des contraintes d'intégrité sur les liens. À savoir, un lien ne peut pas exister par lui-même, il doit avoir une origine et une destination qui doivent être des connaissances abstraites ou factuelles. Un lien peut connecter une connaissance à elle-même (à la fois origine et destination). Une connaissance peut être liée à aucune, une, ou plusieurs connaissances. Entre deux types de connaissance, les seuls liens valides considérés sont donnés dans le Tableau 1. S'il y a un lien entre deux connaissances alors il est unique avec un type unique. S'il y a plusieurs destinations pour un lien S, I/P ou I alors ces destinations doivent avoir le même type (ce qui n'est pas exigé pour les autres liens).

De plus, Paquette (Paquette *et al.*, 2006) a imposé que les relations de Spécialisation (S), Composition (c) et Précédence (P) soient des ordres strict partiel (irréflexifs, transitifs, asymétriques et non totaux) [3] et que les liens Intrant/Produits (I/P), Régulation (R) et Instanciation (I) ne soient pas transitifs.

L'approche MOT est intéressante parce qu'elle a introduit la notion d'éléments et de relations typés, nous allons toutefois exprimer quelques critiques qui justifient la tentative de définir un nouveau langage.

3. Par soucis de simplicité, les liens transitifs ne sont pas exprimés dans les schémas.

*Tableau 1 – Liens valides entre connaissances*

| | Connaissance Abstraite | | | Connaissance Factuelle | | |
|---|---|---|---|---|---|---|
| *De* \ *À* | **Concept** | **Procédure** | **Principe** | **Exemple** | **Trace** | **Enoncé** |
| **Concept** | S, I, C | I/P | R | I, C | I/P | R |
| **Procédure** | I/P | C, S, P, I | C, P, R | I/P | I, C, P | R, P, C |
| **Principe** | R | C, R, P | C, S, P, R, I | R | C, R, P | I, C, P |
| **Exemple** | C | I/P | R | C | I/P | R |
| **Trace** | I/P | P, C | P, R, C | I/P | C, P | C, P, R |
| **Énoncé** | R | R | R | R | C,R,P | C,R,P |

### 3.2 Critiques

Un des objectifs de MOT était d'être facile à apprendre et à comprendre. À notre avis, ce but n'est pas atteint. Les lettres employées pour différencier les différents types de liens ne sont pas très explicites, il serait plus faciles d'employer des formes, des couleurs, des mots, des symboles ou des émoticônes... Les formes standard de MOT n'ont pas de signification intrinsèque, par exemple, les procédures ne sont pas modélisées en prenant en compte la notion temporelle/causale forte qui les caractérisent (elle pourrait être capturée par une roue dentée ou une flèche...). Par ailleurs, les principes n'ont pas de définition très cadrée, par conséquent ils sont souvent représentés par des phrases complètes, ce qui va à l'encontre de l'idée d'employer des modèles visuels privilégiant la simplicité et la clarté. Par conséquent le postulat *Brièveté, puissance suggestive et clarté des symboles* n'est pas satisfait par MOT.

Des liens de Régulation semblent être employés d'une manière inexacte dans l'exemple fourni par l'auteur : un lien de Régulation est employé pour exprimer la manière (par exemple « burn with an owen » dans la figure 2) ou pour exprimer des liens de priorité entre les concepts, cette utilisation n'est ni claire ni univoque violant les postulats de *Correspondance de similarités* et *Significativité*.

D'autre part, les diagrammes complexes peuvent être difficiles à appréhender, violant le principe d'*Efficacité d'accès* et de *Limitation de la surcharge cognitive*. La méthode MOT n'offre pas la possibilité d'effectuer des zoom avant/arrière afin de faire une projection selon certaines relations d'intérêt ou certains attributs précis : la géographie, la causalité, la spécificité, etc. Par exemple, les relations de Composition et d'Instances sont des relations qui changent le niveau de détails, par conséquent, elles pourraient être associées à certaines opérations de zoom avant ou arrière. Ceci viole le postulat *3D*.

## 4 Notre projet : VTL

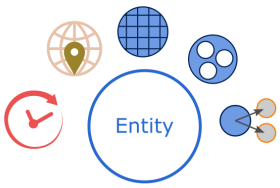
Comme dans MOT, nous proposons d'utiliser trois types d'éléments : actions, entités et conditions. Notre amélioration porte plutôt sur les relations entre ces différents éléments.

### 4.1 Les 3 types d'éléments

Les trois types d'éléments peuvent être génériques (bordure bleue et fond blanc) ou spécifiques (bordure orange et fond gris). Chaque élément générique peut être accompagné du symbole *Instances*.

La fonctionnalité *Instances* décrit des exemples particulier d'une entité générique. Pour accéder aux entités génériques auquel l'entité spécifique courante se réfère on utilise un lien hypertexte. Par exemple, «  four n$^o$ref F118 » est une instance de « four ».

### 4.1.1 Entités




Chaque entité (générique ou spécifique) est représentée par un cercle, avec un intitulé et une icône (optionnelle) à l'intérieur. Cette forme est associée à cinq symboles. Les cinq symboles dessinés ci-dessus, en commençant de la droite vers la gauche représentent respectivement les caractéristiques *Instances* ou *Instance de* (selon que l'entité est générique ou spécifique), *Composé de*, *Propriétés*, *Spatialisation* et *Temps*. Cela permet d'associer les entités à certaines spécificités et permet en outre de naviguer par projection sur une caractéristique spécifique.

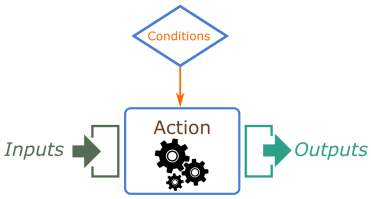
La fonctionnalité *Composé de* peut décrire un ensemble d'entités qui composent l'entité principale. Par exemple, une voiture est composée d'un volant, de quatre roues, d'un moteur, etc.

La fonctionnalité *Propriétés* peut décrire des caractéristiques spécifiques de l'entité. Par exemple, un « déchet » peut être « macroscopique », ou « jaune » ou « alimentaire ».

La fonctionnalité *Spatialisation* est généralement utilisée pour localiser l'entité dans le monde, dans une pièce, etc. Par exemple, il est possible de spécifier les coordonnées GPS, les volumes et les zones, ainsi que les positions relatives (au-dessous/ au-dessus /à gauche /à droite d'une autre entité).

La caractéristique *Temps* quant à elle permet de localiser l'entité dans un temps universel ou relatif (c'est-à-dire, par rapport à d'autres entités). Par exemple, une voiture peut avoir une date de naissance, et une durée de vie moyenne.

### 4.1.2 Actions




Les actions sont symbolisées par un rectangle bleu avec une étiquette de couleur brune associée au dessin d'un engrenage. Les actions sont liées à des entrées, des sorties et des conditions.

Les entrées sont les entités nécessaires pour exécuter l'action, ou les entités qui sont intéressantes à mentionner pour que l'action ait lieu. Par exemple « déchets » et « four » sont des entrées pour l'action « brûler des déchets ».

Les sorties sont des entités résultant d'une action ou qui ont un certain intérêt à être mentionnées après une action. Par exemple, l'action « brûler des déchets » est liée à un ensemble de sortie telles que « résidus », « gaz » et « four ».

Les conditions sont décrites dans la section suivante.

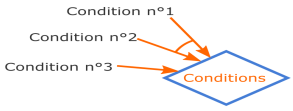
### 4.1.3 Conditions


Les conditions sont représentées par des losanges bleus avec un libellé orange. Certaines actions nécessitent des conditions pour se réaliser.

Dans notre exemple de la figure 4 : le four doit être en état de marche.

Plus généralement, les conditions peuvent s'exprimer sur toutes les entités, même si elles ne sont pas liées à des actions. Plusieurs conditions peuvent être connectées par des opérateurs logiques et / ou. Nous utilisons la convention des *arbres et-ou* pour représenter des combinaisons de conditions.

Par exemple, la condition définie par ((condition n$^o$1) *et* (condition n$^o$2)) *ou* (condition n$^o$3) est représentée par :




## 4.2 Relations

Dans VTL, les relations entre entités, actions et conditions sont représentées au moyen des 5 caractéristiques prédéfinies associées aux entités, ou par les 3 caractéristiques associées aux actions ou encore par une combinaison de conditions. Cela nous permet de mettre en avant les relations suivantes.

- Relations d'entités à entités : comme par exemple *Composé de*, *Instances*, *Forme générique de*, *Se déroule avant*, *Se déroule après*, *Au nord de*, etc.
- Relations entre entités et actions : via les entités entrée/sortie propres aux actions.
- Relations entre conditions et entités : certaines conditions peuvent être exprimées sur les entités. Elles peuvent être considérées comme des filtres sur ces entités.
- Relations entre conditions et actions : ces relations sont des contraintes sur l'exécution possible d'une action.
- Relations entre conditions et conditions : les relations entre les conditions sont symbolisées par la convention et-ou (comme on l'a vu dans la section 4.1.3)

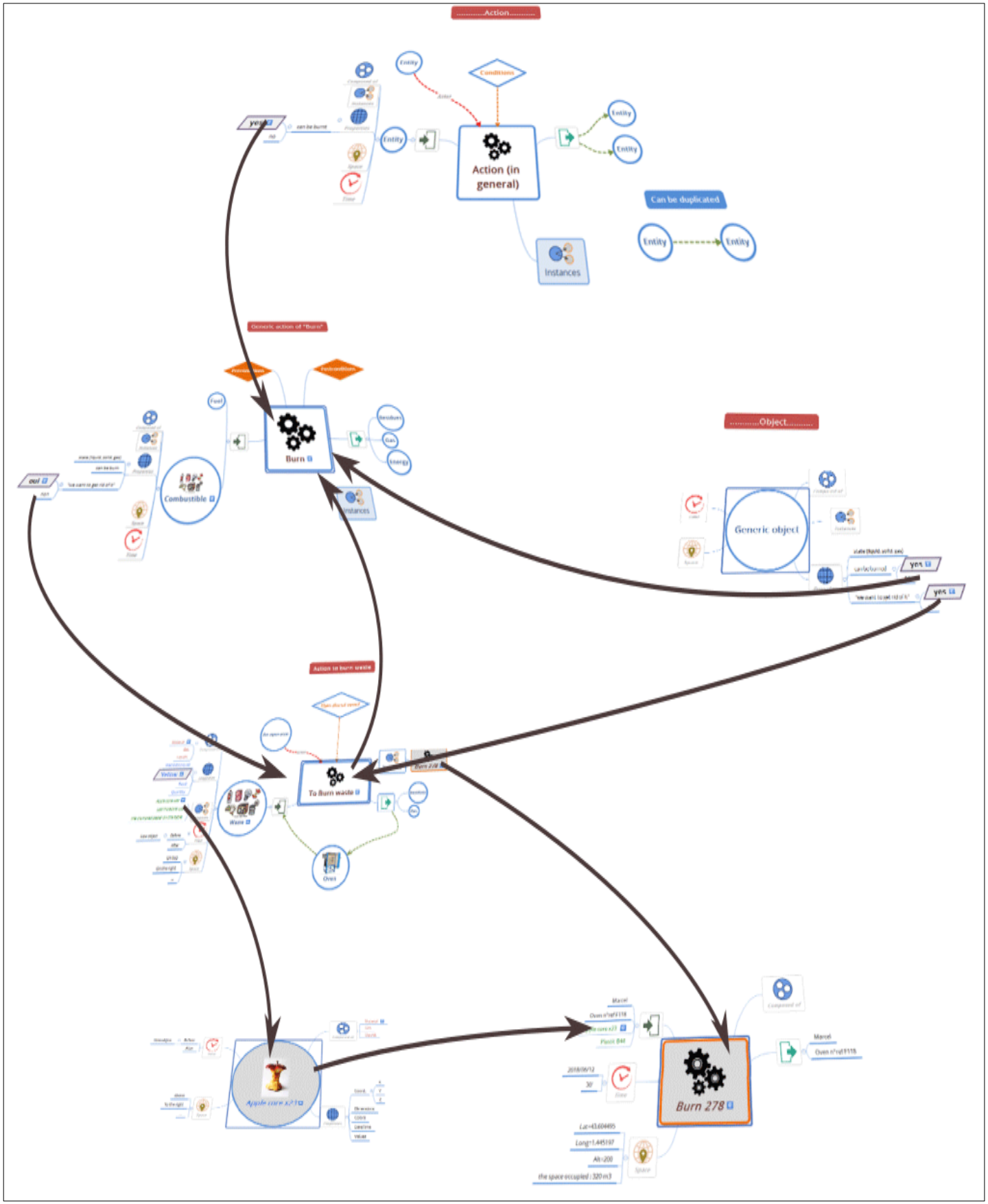
## 4.3 Navigation

VTL est associé à un outil de navigation. En effet, afin de permettre une représentation plus claire, au lieu d'avoir un vaste graphique d'entités, il est possible de ne représenter qu'un sujet principal sans trop de détails. Ensuite, le navigateur peut naviguer par zoom avant/arrière sur une caractéristique précise pour n'obtenir que les détails qui l'intéressent. En cliquant sur une entité, l'utilisateur obtient une nouvelle vue dont l'entité occupe le centre. La navigation est ainsi un moyen d'explorer certaines caractéristiques précises. La navigation est illustrée par la Figure 3 disponible en ligne (Dupin de Saint Cyr & Parade, 2018).

## 4.4 Exemple

La figure 4 décrit une représentation en VTL de la gestion des déchets décrite en MOT par les figures 1 et 2. Cette figure représente l'action générique de brûler les déchets et une instance de cette action peut être examiné en cliquant sur « Brûler-278 » à partir de l'attribut *Instances* de l'action principale «  Brûler des déchets » (Figure 5).

FIGURE *3 – Navigation en VTL*



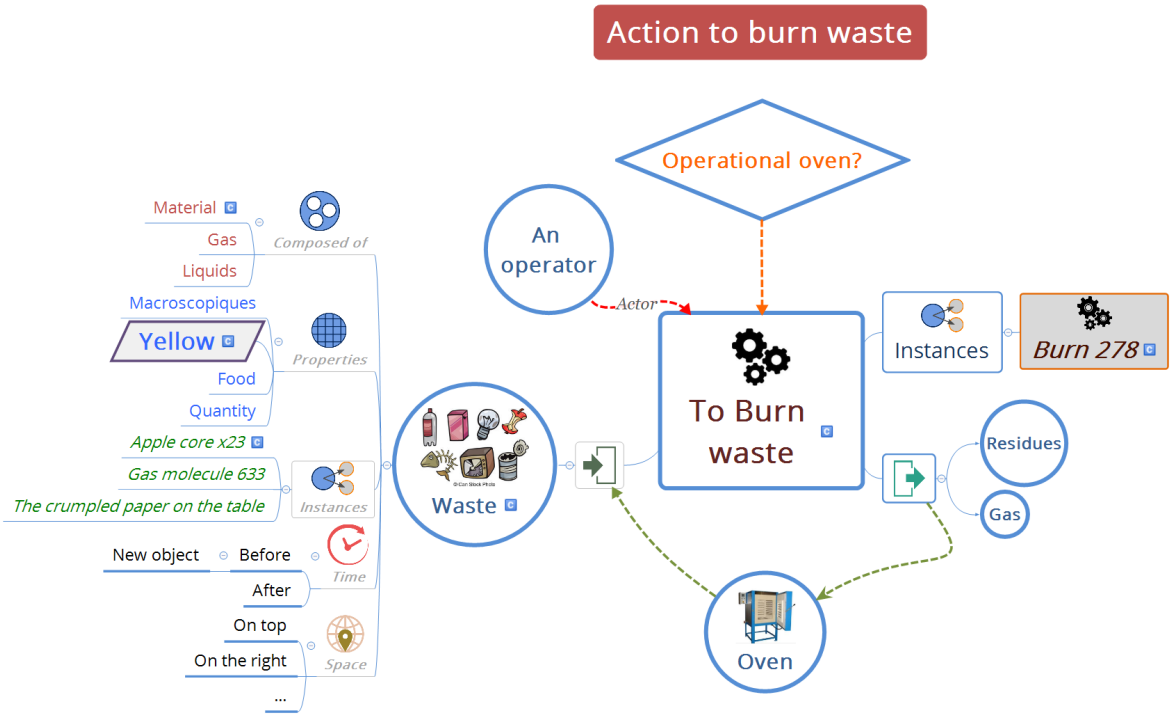



FIGURE 4 – *L'action « To Burn Waste » en VTL*

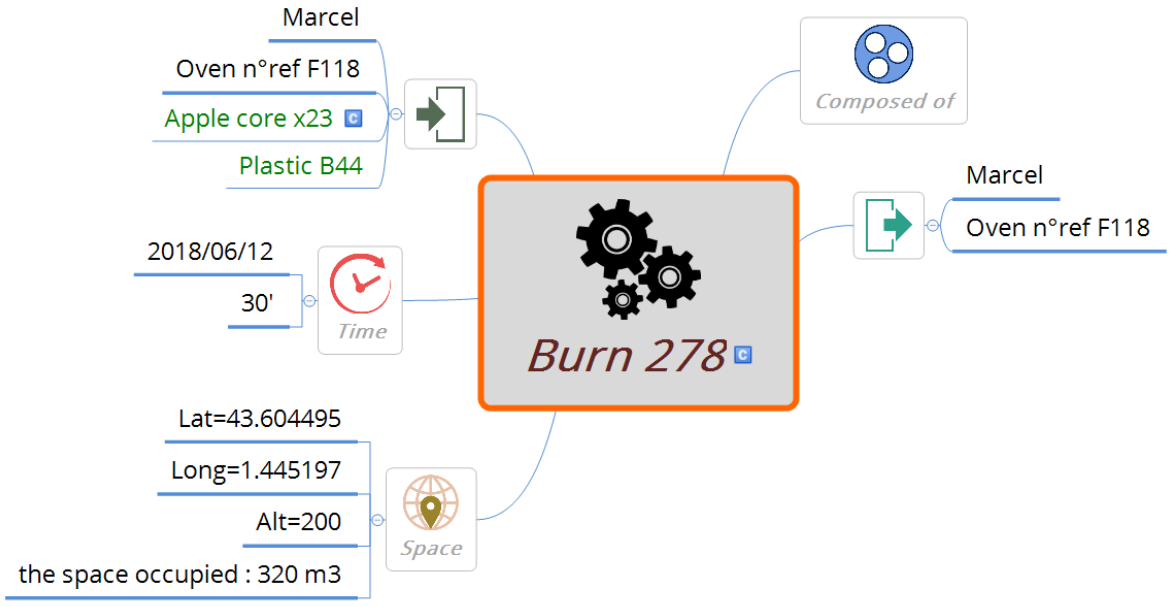



FIGURE 5 – *Instance de l'action Burn en VTL*

## 5 Propriétés de VTL

### 5.1 Propriétés sémiotiques

Comme mentionné dans l'introduction, nous nous appuyons sur les travaux et recherche liés à la sémiologie. Ceux de *Jacques Bertin* (Bertin, 1973), cartographe français, a ainsi mis en avant dans son ouvrage majeur « Sémiologie graphique » un ensemble de 8 « variables visuelles ». Ces travaux étant plutôt dédiés à la représentation des données, nous ne tiendrons compte que de quelques-unes de ces variables visuelles : comme le changement de couleur pour signifier un changement de catégories (« entité générique » ou « instance » pour nous), de formes (« entités » / « actions » / « conditions » dans notre cas) et de taille de ces dernières. En l'état actuel de nos travaux, ce dernier élément n'est pas complètement implémenté. Nous excluons les autres variables visuelles (x, y [4], grain, valeur d'une couleur, orientation) principalement utilisées dans les cartes géographiques.

La théorie de la *Gestalt* (« théorie de la forme ») est également une source d'inspiration pour nous. Elle fait référence à un ensemble de lois que la perception visuelle humaine met en œuvre de façon automatique. La loi de proximité par exemple, qui crée implicitement une notion de groupement d'entités proches les unes des autres, sera abondamment utilisée. De même, le principe de continuité, permettra dans VTL de guider l'œil du lecteur pour l'amener « naturellement » au symbole visuel suivant, sans rupture (tel que celui généré par un saut de page par exemple). L'utilisation de lien hypertexte dans la version actuelle du prototype de VTL est cependant en contradiction avec ce principe. Nous envisageons de corriger cet inconvénient dans la version suivante de VTL et d'ajouter le principe de *Continuité* à nos axiomes.

**Comment avons-nous choisi les symboles visuels de VTL ?**

Les *entités* et les *attributs* associés sont tous dans des *cercles*. Cette forme est très souvent utilisée pour symboliser quelque chose d'indéfini, de non précisé. On la retrouve dans de nombreux logiciels pour représenter un objet, une personne, une abstraction, etc.

Les symboles choisis pour les attributs sont décrits ci-dessous.

 La forme *Composé de* montre un cercle qui en englobe d'autres. On visualise implicitement une structure arborescente qui correspond bien au « tout » composé de « parties » ;

 Le symbole *Propriétés* affiche quant à lui une grille, un filtre : parmi toutes les propriétés possibles de l'entité en question, seules quelques-unes sont utilisées, filtrées, sélectionnées ;

 Le symbole *Instance* utilise la forme des entités, dont l'une est la "source" (en bleu) et les deux autres "instanciées" (en gris). Le changement de couleur est nécessaire afin d'indiquer qu'il ne s'agit pas de la même nature d'objets ;

 *Spatialisation* fait référence à une mappemonde et à marqueur proche de celui utilisé dans Google Maps, deux symboles qui ne laissent aucune ambiguïté sur la localisation dans l'espace de l'entité ;

 Le symbole *Temps* combine également 2 références visuelles : les aiguilles d'une horloge (référence au temps) au centre d'un cadran fléché pour signifier la notion de durée.

Le symbole associé à une *Action* comporte un engrenage qui évoque un mécanisme en mouvement. En accord avec les variables visuelles de J. Bertin, nous utilisons une forme

4. La position des éléments dans la représentation VTL est cependant prise en compte mais pas de manière absolue.

différente de celles des entités, le carré est choisi car c'est également cette forme que l'on retrouve dans la cartographie de processus.

 Les entrées et sorties sont représentées par des portes.

La forme des *Conditions* reprend celle utilisée dans les logigrammes pour tester une variable et aiguiller le déroulement du procédé.

Conscient que ces choix sont subjectifs nous envisageons de développer un protocole expérimental de validation par des utilisateurs du langage VTL et plus précisément des symboles utilisés. Au-delà de cette validation expérimentale, il serait intéressant d'étudier la possibilité de mettre en place des principes formels sémiotiques (comme le principe de continuité). Un autre principe sémiotique pourrait être mis en place pour les contraires : il faudrait imposer une cohérence en couleur/forme/taille entre contraire. Ici, par exemple, les couleurs des portes Entrée et Sortie sont différentes mais les couleurs pas significative de la fonction contraire. Cependant ce n'est pas encore l'objet de notre présente contribution. Notons que cette piste a été explorée pour le langage VCM (Visualization of Concepts in Medicine (Lamy *et al.*, 2013)). Le langage VCM propose une liste de symboles de base représentant des concepts médicaux que l'utilisateur peut combiner pour former de nouveaux symboles. Cela permet d'accompagner les textes médicaux avec ces images, chaque image est facile à comprendre à partir de sa décomposition en symboles de base. L'utilisation de VCM satisfait ainsi notre principe de *Brièveté, puissance suggestive et clarté des symboles*. La démonstration du bien fondé du langage VCM donnée dans (Lamy *et al.*, 2013) se base sur l'association d'une ontologie OWL aux symboles dont la traduction en DL a permis de prouver la cohérence hiérarchique des icones.

## 5.2 Propriétés théoriques

Il est possible d'imposer dans VTL que tout schéma n'ait qu'une seule entrée qui serait le centre d'un graphe, ce qui impliquerait que les principes *Accessibilité*, *Navigation*, *Efficacité de l'accès* et *Entrée* seraient respectés. Dans VTL, les symboles que nous proposons sont associés à des mots courts et nous recommandons d'ajouter des images avec ces éléments. Par conséquent, nous incitons à utiliser le « double codage » qui implique la satisfaction du principe de *Brièveté, puissance suggestive et clarté des symboles*. Le double codage introduit par (Paivio, 1990) consiste à

Par construction, les positions/formes/couleurs des trois types d'éléments ont été choisis dans VTL afin d'avoir une interprétation claire, ce qui signifie que le principe de *Correspondance de similarité* est respecté pour les formes et les couleurs, ainsi que le principe de *Significativité*.

Concernant les positions relatives des objets, leur *Significativité* n'est pour le moment imposé que pour les actions (où les entrées sont sur la gauche tandis que les sorties sont à droite, et les conditions sont au-dessus).

La navigation dans VTL est faite de manière à donner la possibilité de zoomer en avant/en arrière (voir Figure 3), ce qui permet de satisfaire le principe *3D*. Le principe de *Limitation de la surcharge cognitive* est un postulat qui devra être imposé aux auteurs de VTL. Cette condition peut être imposée sans perte d'information en utilisant les fonctionnalités natives de VTL (navigation et zoom avant/arrière).

Pour illustrer notre propos, nous avons utilisé le logiciel XMind qui nous a permis de créer très facilement un exemple complet. D'où le postulat sur la *Facilité d'écriture* même si un outil dédié serait plus adapté. Concernant le postulat *Traduisible*, l'idée d'utiliser un langage typé nous permet d'imposer des restrictions sur les types d'éléments autorisés et également sur leurs connexions. Une traduction automatique de toute expression sera le sujet d'une prochaine étude. Le principe de *Cohérence vérifiable* n'est pas encore en œuvre en VTL, il sera possible lorsque nous disposerons d'un outil de traduction formelle.

## 6 Conclusion et travaux connexes

Nous avons proposé plusieurs principes visant à régir les langages de représentation visuelle des connaissances et avons défini un nouveau langage de représentation typé VTL compatible avec ces principes. En VTL l'information est accessible par navigation à l'intérieur d'une structure arborescente. En effet, à un instant donné, l'accès à tous les détails d'un sujet donné est superflu pour l'utilisateur. Au contraire, il est nécessaire de limiter les détails exposés d'emblée afin de ne pas lui imposer une charge mentale trop importante. Ainsi, la navigation dans VTL permet à l'utilisateur d'accéder aux niveaux de détails voulus seulement pour les informations qui l'intéressent spécifiquement.

Comme nous le signalons dans la précédente section, la validation sémiotique de notre langage par des tests avec des utilisateurs n'est pas encore envisagée. Nous n'en sommes pas au stade de la définition d'un protocole expérimental qui pourrait étudier la facilité de passage dans un sens ou dans l'autre d'un texte écrit à sa représentation en VTL. Nous pourrions utiliser ce protocole pour comparer VTL avec d'autres langages visuels standards comme UML.

Le logiciel XMind a été utilisé pour créer des exemples et pour simuler la navigation, mais le développement d'une interface utilisateur graphique (GUI) spécifique à VTL est à l'étude. De plus, notre prochaine étape consistera à étudier la traduction de VTL dans un langage non visuel formel afin de proposer des inférences et des contrôles de cohérence.

Les notions d'héritage de propriétés sont traduites en VTL par navigation vers un niveau plus spécifique ou plus générique. L'héritage de propriétés n'est pas un concept nouveau inventé avec les langages objets puisque d'après Sowa (Sowa, 2006), ça n'est qu'« un cas particulier des syllogismes d'Aristote dans lequel on spécifie les conditions d'héritage de propriétés d'un type vers un sous-type ». De même, Sowa souligne que la distinction entre instances et catégories est déjà présente dans la logique médiévale et se retrouve en bas du dessin de l'arbre de Porphyre, attribué à Pierre d'Espagne, et daté de 1239 (voir Figure 6). Ces notions de relations variées entre concepts associées à des relations d'héritage sont présentes dans les réseaux à héritage de structure (*structure inheritance networks* comme le langage KL-ONE (Knowledge Language One) (Brachman *et al.*, 1991) voir Figure7. Un des points intéressants de l'arbre de Porphyre comme d'UML est l'existence d'attributs appelés *Differentiae* qui permettent de séparer un type d'un sous-type.

Une autre direction de travail consisterait à étudier comment VTL permet d'englober les liens qui ne sont pas traités par MOT, à savoir les relations RCC8 (Randell *et al.*, 1992) ou les intervalles d'Allen, ou d'autres relations entre concepts (on pourrait se référer par exemple à la typologie des mots-liens écrit par Christian Barette (Barette, 2002)).

Dans l'idée de raisonner visuellement, nous envisageons la possibilité de transformer le dessin par inférence, c'est un peu réalisé dans le langage VCM (Lamy *et al.*, 2013) puisqu'il permet de créer des combinaisons de symboles à partir de symboles. Cependant une autre façon de réaliser ce raisonnement est d'écrire des règles sémantiques d'insertion ou de retrait d'éléments graphiques, comme cela a été proposé en 1896 par Charles S Pierce avec les graphes existentiels (*existential graphs*) (Roberts, 1973). Les graphes existentiels sont des représentations dans lesquels on peut entourer des lettres (pour la négation) ou les juxtaposer (pour la conjonction), ce qui permet de visualiser une formule logique, avec des règles disant que deux entourages d'un même élément peuvent être supprimés (double négation). Bien que la représentation proposée ne soit pas très intuitive pour la perception visuelle, la gestion sémantique d'énoncés logique faite par Pierce pourrait nous inspirer afin de mettre en place des règles de modifications (ou d'inférence) en VTL.

## Remerciements

Les auteurs remercient les rapporteurs anonymes du comité de programme de la conférence IC'2018 pour leurs compétences et leur exigence. Leurs remarques pertinentes ont permis de faire progresser cet article de façon substantielle.

FIGURE *6 – L'arbre de Porphyre selon Pierre d'Espagne (1239) (Sowa, 2006)*

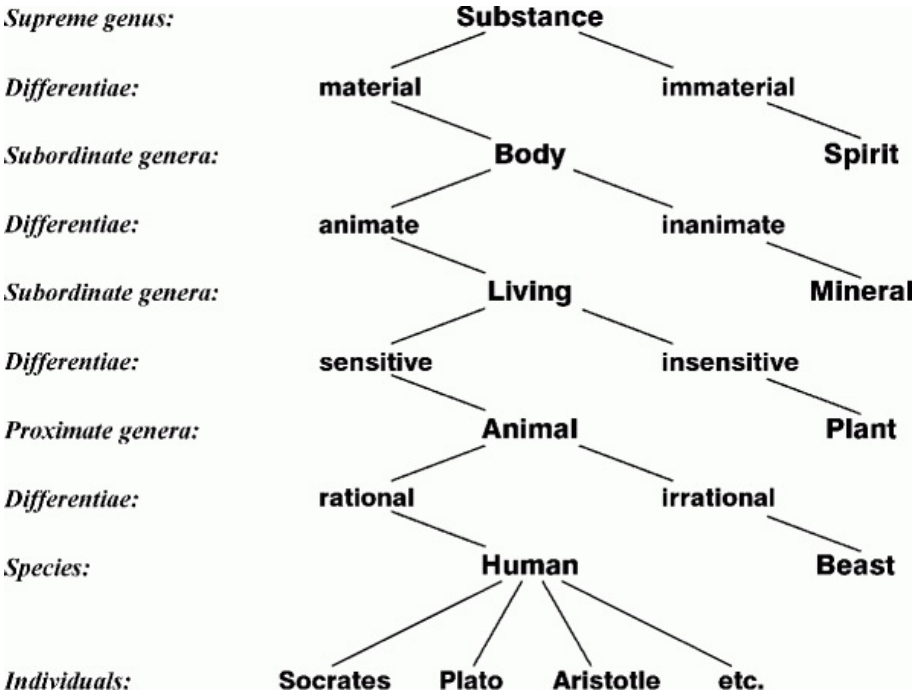
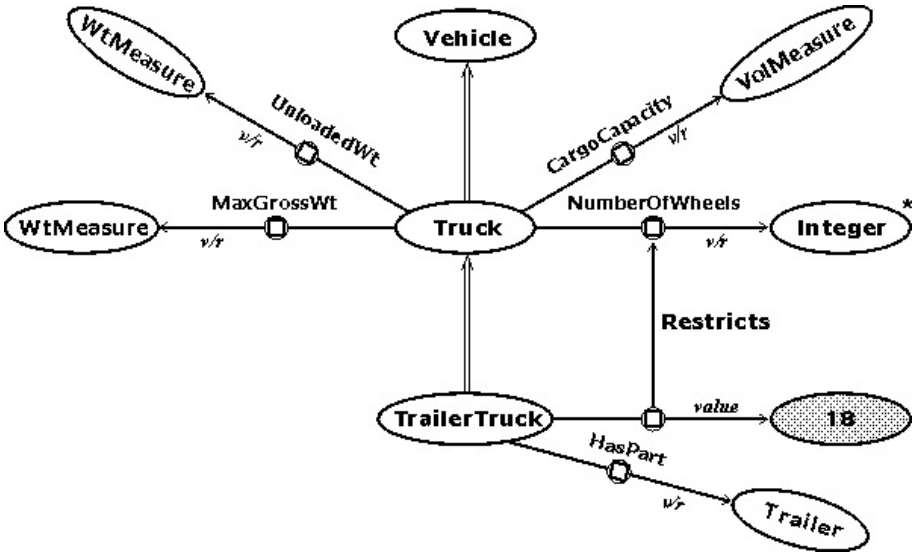



FIGURE *7 – Représentation en KL-ONE d'un camion (Sowa, 2006)*



## Références


BARETTE C. (2002). *L'analyse des relations*. Rapport interne, Université de Sherbrooke, Québec.

BERTIN J. (1973). Sémiologie graphique : Les diagrammes-les réseaux-les cartes. *Gauthier-VillarsMouton & Cie*.

BRACHMAN R. J., MCGUINNESS D. L., PATEL-SCHNEIDER P. F., RESNICK L. A. & BORGIDA A. (1991). Living with classic : When and how to use a kl-one-like language. *Principles of semantic networks*, p. 401–456.

BUZAN T. (1974). *Use your head*. London : BBC Books.

DICKOVER M. E., MCGOWAN C. L. & ROSS D. T. (1977). Software design using : Sadt. In *Proceedings of the 1977 annual conference*, p. 125–133 : ACM.

D'SOUZA D. F. & WILLS A. C. (1998). *Objects, components, and frameworks with UML : the catalysis approach*, volume 1. Addison-Wesley Reading.

DUPIN DE SAINT CYR F. & PARADE D. (2018). Knowledge Representation in a Visual Typed Language : from Principles to Practice. In *International Conference on Information Visualisation (IV)*, KV : Knowledge Visualization and Visual Thinking : IEEE.

DUPIN DE SAINT CYR F. & PARADE D. (2018). Maquette interactive VTL sur l'exemple Burn Waste. `http://www.scenario-interactif.com/Maquette-VTL.xmind`.

GRUNDSTEIN M. (2002). Gameth : un cadre directeur pour repérer les connaissances cruciales pour l'entreprise. *Lamsade Université Paris-Dauphine*.

GRUNDSTEIN M. & ROSENTHAL-SABROUX C. (2004). Gameth®, a decision support approach to identify and locate potential crucial knowledge. In *Proc. 5th Europ. Conf. on Knowledge Management*, p. 391–402.

ISO (1985). 5807 : 1985 information processing-documentation symbols and conventions for data, program and system flowcharts, program network charts and system resources charts. *Geneva*.

LAMY J.-B., SOUALMIA L. F., KERDELHUÉ G., VENOT A. & DUCLOS C. (2013). Validating the semantics of a medical iconic language using ontological reasoning. *Journal of Biomedical Informatics*, **46**(1), 56 – 67.

MARCA D. A. & MCGOWAN C. L. (1987). *SADT : structured analysis and design technique*. McGraw-Hill, Inc.

MILLER G. A. (1956). The magical number seven, plus or minus two : Some limits on our capacity for processing information. *Psychological Review*, **63**, 81–97.

NEURATH O. (1936). International picture language. the first rules of isotype. *London : K. Paul, Trench, Trubner & Co*.

NOVAK J. D. & CAÑAS A. J. (2008). *The theory underlying concept maps and how to construct and use them*. Rapport interne, Institute for Human and Machine Cognition.

PAIVIO A. (1969). Mental imagery in associative learning and memory. *Psychological Review*, **76**, 241-263.

PAIVIO A. (1990). *Mental representations : A dual coding approach*. Oxford University Press.

PAQUETTE G. (2010). Visual knowledge and competency modeling-from informal learning models to semantic web ontologies. *Hershey, PA : IGI Global*.

PAQUETTE G., LÉONARD M., LUNDGREN-CAYROL K., MIHAILA S. & GAREAU D. (2006). Learning design based on graphical knowledge-modelling. *J. Educational Technology & Society*, **9**(1).

PEPPER S. (2000). The TAO of topic maps. `https://www.researchgate.net/publication/225070304_The_TAO_of_topic_maps`, p. 1–21.

RANDELL D. A., CUI Z. & COHN A. G. (1992). A spatial logic based on regions and connection. In *Proc. of the 3rd Int. Conf. on Principles of Knowledge Representation and Reasoning*, p. 165–176.

ROBERTS D. D. (1973). *The existential graphs of Charles S. Peirce*, volume 27. Walter de Gruyter.

RUMBAUGH J., BLAHA M., PREMERLANI W., EDDY F. & LORENSEN W. E. (1991). *Object-oriented modeling and design*, volume 199, 1. Prentice-hall Englewood Cliffs, NJ.

RUMBAUGH J., BOOCH G. & JACOBSON I. (2017). *The unified modeling language reference manual*. Addison Wesley.

SOWA J. (1984). *Conceptual structures—Information processing in mind and machine*. MA : Addison-Wesley, Reading.

SOWA J. F. (2006). Semantic networks. *Encyclopedia of Cognitive Science*.

SWELLER J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, **4**(4), 295–312.

TEBOUL J. & DAMIER P. (2017). *Neuroleadership*. Odile Jacob.

WIKIPÉDIA (2017). Modélisation par objets typés.

# Construction collaborative d'ontologies qui intègre l'utilisateur et son contexte

La définition d'ontologie la plus souvent reprise est celle de Gruber : "une spécification explicite d'une conceptualisation *partagée*". Cette session se focalise sur cette notion de *partage* qui est centrale et donc sur les utilisateurs et leur contexte. La première contribution étudie l'évolution du modèle de l'utilisateur de systèmes de construction collaborative d'ontologies. La deuxième contribution choisit d'utiliser des traces évolutives pour caractériser les utilisateurs. Enfin, parce que la collaboration ne va pas forcément de soi, la dernière contribution vise à prendre en compte différents critères pour la rendre efficace.

# Étude de l'évolution du modèle de l'utilisateur des systèmes de construction collaborative d'ontologies

Alain Giboin

UCA Inria, CNRS, I3S
Équipe Wimmics
2004 route des Lucioles
06992, Sophia Antipolis, France
alain.giboin@inria.fr

**Résumé** : Cet article rend compte d'une étude en cours sur l'évolution du modèle de l'utilisateur de systèmes de construction collaborative d'ontologies. Par *modèle de l'utilisateur* (ou *modèle du contributeur*), nous entendons la représentation que les concepteurs se font des utilisateurs de leurs systèmes et plus généralement des contributeurs à la construction des ontologies. Nous décrivons : 1) la méthode que nous utilisons pour étudier l'évolution du modèle de l'utilisateur ; 2) l'évolution de ce modèle (en termes de types d'utilisateurs, de caractérisations de l'utilisateur et de caractérisations de l'environnement de l'utilisateur) ; 3) les évolutions parallèles : a) des méthodes de conception des systèmes collaboratifs ; b) des systèmes eux-mêmes ; et c) des méthodes de construction collaborative des ontologies. Nous mentionnons quelques perspectives d'évolution envisagées par les concepteurs eux-mêmes. Cette étude vise à faire ressortir l'importance d'acquérir une meilleure connaissance des contributeurs potentiels à la construction collaborative des ontologies afin d'obtenir des outils collaboratifs mieux adaptés à ces contributeurs.

**Mots-clés** : Systèmes de construction collaborative des ontologies ; Ingénierie ontologique participative Méthodes pour l'ingénierie ontologique ; Modèle de l'utilisateur ; Modèle du contributeur ; Evolution des modèles.

## 1 Introduction

Dans le domaine de la conception et de l'évaluation de l'Interaction Humain-Machine (IHM), on désigne par *modèle de l'utilisateur* : ou bien *1)* le *modèle-de-l'utilisateur du concepteur* (la représentation que le concepteur se fait de l'utilisateur de son système) ; ou bien *2)* le *modèle-de-l'utilisateur du système* ou *modèle embarqué dans le système* (la représentation que « se fait » le système de son utilisateur au fur et à mesure de son interaction avec ce dernier) ; ou bien *3)* le *modèle-mental-du-système de l'utilisateur* (la représentation que l'utilisateur se fait du système) (cf. Kelly & Colgan, 1992).

On s'intéresse ici au premier type de modèle et, plus précisément, au *modèle-de-l'utilisateur des concepteurs de systèmes de construction collaborative d'ontologies*, ou *modèle-du-contributeur* de ces concepteurs (les utilisateurs contribuant d'une façon ou une autre à la construction des ontologies). Cet article rend compte d'une étude en cours sur l'évolution de ce modèle entre le début des années 1990 et aujourd'hui, où l'on verra qu'initialement destinés à des groupes d'ontologues ou d'ingénieurs de la connaissance, les systèmes de construction collaborative d'ontologies se sont ouverts progressivement à d'autres types d'utilisateurs/contributeurs.

Par cette étude nous souhaitons montrer qu'une meilleure connaissance des utilisateurs/contributeurs – reflétée par un modèle de l'utilisateur/contributeur plus réaliste –

est importante pour concevoir des systèmes adaptés. Cette étude se veut aussi une invitation à pratiquer plus avant l'analyse et la modélisation de l'utilisateur/contributeur de systèmes de construction collaborative d'ontologies.

Dans la suite de cet article, nous fournissons une première description : *1)* de la méthode que nous utilisons pour étudier l'évolution du modèle de l'utilisateur/contributeur ; *2)* de l'évolution de ce modèle (en termes de types d'utilisateurs, de caractérisations de l'utilisateur et de caractérisations de l'environnement de l'utilisateur) ; *3)* des évolutions parallèles : a) des méthodes de conception des systèmes de construction collaborative d'ontologies ; b) des systèmes eux-mêmes ; c) des méthodes de construction collaborative des ontologies et d) des perspectives d'évolution. Nous mentionnons pour terminer quelques perspectives d'évolution envisagées par les concepteurs eux-mêmes.

## 2 Méthode d'étude de l'évolution du modèle utilisateur

Dans cette section nous décrivons la méthode que nous utilisons pour étudier l'évolution du modèle de l'utilisateur des systèmes de construction collaborative d'ontologies.

### 2.1 Sources consultées

Nous consultons principalement les types suivants de sources : *1)* des travaux décrivant les systèmes de construction collaborative d'ontologies ;qu'ils s'appellent {éditeurs | outils | boîtes à outils | plates-formes | environnements | …} de {construction | conception | ingénierie | …} d'ontologies outils, boites à outils ou autre (voir Table 1) ;*.2)* des travaux décrivant des systèmes spécifiques et leurs utilisateurs (voir table 1, une liste non exhaustive de systèmes) ; *3)* des états de l'art sur les systèmes (ex. : Simperl & Luczak-Rösch, 2014 ; Correndo & Alani, 2007) ou des articles comportant un état de l'art détaillé (ex. : Mangione et al., 2011) ; *4)* des travaux décrivant des méthodes de construction collaborative d'ontologies supportées par un système ; *5)* des travaux rapportant des études d'utilisation des systèmes de construction collaborative d'ontologies.

*TABLE 1 – Systèmes de construction collaborative d'ontologies recensés (liste non exhaustive).*

| **Nom** | **Description** |
|---|---|
| APECKS | APECKS = Adaptive Presentation Environment for Collaboration Knowledge Structuring<br>A tool for ontology construction with internal and external KA support<br>*Destiné aux experts du domaine*<br>(Tennison et al., 2002) |
| CoB Editor<br>ATO Editor | COB = Collaborative Ontology Building<br>An ontology editing tool which exploits the notion of modular ontologies to support sharing, reuse, and collaborative editing of partial order (i.e., DAG-structured) ontologies<br>(Bao et al., 2006) |
| | ATO = Animal Trait Ontologies<br>Cet édieur est une adaptation de l'éditeur CoB<br>(Bao et al., 2006) |
| COE | COE = Collaborative Ontology Editor<br>(Xexéo et al., 2005) |
| CODE-COE | COE = Collaborative Ontology Environment<br>Based on CMapTools<br>*Tool dedicated to domain experts*<br>(Hayes et al., 2003, 2005) |
| Collaborative Protégé<br>Web Protégé<br>uComp Protégé Plugin | Collaborative Protégé : An extension of the existing Protégé system that supports collaborative ontology editing<br>(Tudorache & Noy, 2007)<br>(Tudorache et al., 2008) |

| | |
|---|---|
| | Web Protégé: a collaborative ontology editor and knowledge acquisition tool for the Web<br>(Tudorache et al. 2013) |
| | uComp Protégé Plugin: A plug-in aimed at crowd-based ontology engineering<br>(Wohlgenannt et al., 2016) |
| Compendium | A tool for augmenting design deliberation in Collaborative Ontology Design<br>(BuckinghamShum et al., 2002) |
| ContentCVS | A tool for supporting concurrent ontology development<br>(Jiménez Ruiz et al., 2010° |
| ECCO<br>Folkon | Editeur collaboratif et contextuel d'ontologies<br>(Giboin & Durville, 2007; Giboin et al., 2008, 2013 ; Corby & Durville, 2009) |
| | Folkon : Editeur de folksonomies<br>Adil El Ghali (non publié) |
| HCONE et SharedHCONE | Human-centered ontology management environments supporting the HCOME methodology<br>(Kotis et al., 2005, 2006) |
| Hozo | A tool for distributed and collaborative construction of ontologies<br>(Kozaki et al., 2007) |
| KPI Onto Editor | Collaborative building of an ontology of key performance indicators (KPI)<br>(Diamantini et al., 2014) |
| MarcOnt Portal<br>WMap Portal | MarcOnt Portal : Outil collaboratif de développement et de gestion des ontologies<br>(Dabrovski et al., 2007) |
| | WMap Portal (variante de MarcOntPortal) : outil collaboratif de mise en correspondance d'ontologies<br>(Dabrovski et al., 2007) |
| MoKi | A collaborative MediaWiki-based tool for modeling ontological and procedural knowledge in an integrated manner<br>(Ghidini et al., 2010) |
| MyOntology | A wiki-based ontology engineering system marrying ontology engineering and collective intelligence<br>(Siorpaes & Hepp, 2007) |
| NeOn Toolkit | An open-source, multiplatform ontology engineering environment, which provides comprehensive support for the ontology engineering life cycle of networked ontologies<br>(Erdmann & Waterfeld, 2012) |
| ONKI | A Tool for Collaborative Ontology Development for the Semantic Web<br>(Valo et al., 2005) |
| [OntoCommand] | A collaborative ontology construction tool with conflicts Detection<br>(Chen et al., 2008) |
| OntoEdit | An collaborative ontology editor that integrates numerous aspects of ontology engineering<br>(Sure et al., 2002) |
| Ontolingua | A tool for collaborative ontology construction<br>(Farquhar et al., 1997) |
| OntoVerse | An ontology wiki supporting all phases of collaborative ontology engineering.<br>(Mainz et al., 2008) |
| OntoWiki | A tool to develop ontologies collaboratively<br>The tool converts human readable maps into a machine readable ontology<br>*Tool dedicated to domain experts*<br>(Auer et al., 2006) |
| Plug-ins collaboratifs associés à Terminae | Outil de collaboration ajouté à *Terminae*, l'outil de construction de ressources termino-ontologiques (RTO)*(en projet)*<br>(Ressad-Bouidghaghen et al., 2013) |
| Semantic Media Wiki | A. semantic wiki engine that supports collaborative ontology development.<br>Mezghani et al., 2016) |
| Soboleo | A tool combining social bookmarking and lightweight engineering of ontologies<br>(Zacharias & Braun, 2007) |
| Tadzebao et WebOnto | Discussing, browsing and editing ontologies on the Web<br>(Domingue, 1998) |

| | |
|---|---|
| Towntology | Un outil d'aide à la construction d'ontologies pré-consensuelles<br>(Keita et al., 2006) |
| TurtleEditor | An ontology-aware Web editor for collaborative ontology development<br>(Petersen et al., 2016) |
| UbisEditor 3 | A tool for collaborative ontology development on the Web<br>(Loskyll et al., 2009) |
| Usable Ontology | Environnement de construction et d'évaluation d'une ontologie<br>(Missikof et al., 2002) |
| VocBench<br>VocBench 2<br>VocBench 3 | A Web application for collaborative development of multilingual thesauri and ontologies complying with Semantic Web standards (SKOS, OWL)<br>(Stellato et al., 2015, 2017)<br>(autre nom pour VocBench : ACSW - Agrovoc Concept Server Workbench) |
| WebODE | An integrated workbench for ontology representation, reasoning, and exchange<br>(Corcho et al., 2002) |
| Wiki@nt | Environnement multi-agents de construction collaborative d'ontologies<br>(Bao & Honavar, 2004) |

### 2.2 Analyse des sources

Dans ces différentes sources, nous cherchons à repérer plusieurs indicateurs de l'évolution du modèle de l'utilisateur des systèmes de construction collaborative d'ontologies, ainsi que des indicateurs des évolutions parallèles des systèmes, des méthodes de conception de ces systèmes, des méthodes de construction collaborative des ontologies et des perspectives d'évolution. Ces indicateurs sont principalement : *1)* les termes désignant les types d'utilisateurs des COEs ; *2)* les caractérisations de ces utilisateurs ; *3)* les caractérisations de l'environnement de l'utilisateur/contributeur ; *4)* la méthode d'analyse et de modélisation de l'utilisateur de COEs ; *5)* le format de représentation du modèle de l'utilisateur ; les fonctions des COEs découlant des caractérisations des utilisateurs ; *6)* la méthode de construction des ontologies en rapport avec les caractérisations des utilisateurs (recherche de correspondances entre éléments de la méthode et caractéristiques des utilisateurs) ; *7)* les perspectives d'évolution – ce que les concepteurs envisagent de réaliser pour adapter davantage leurs systèmes aux utilisateurs ; *8)* la méthode d'évaluation du COE utilisée : évaluation avec ou sans des utilisateurs (cf. inspection de l'interface) (quelle représentation a-t-on de l'utilisateur lors de cette phase du cycle de conception du système).

## 3 Évolution du modèle de l'utilisateur/contributeur

Dans cette section, nous décrivons plusieurs directions d'évolution du modèle de l'utilisateur/contributeur que nous avons repérées. Nous avons regroupé ces directions en trois catégories : *1)* évolution des types d'utilisateurs / contributeurs impliqués ; *2)* évolution des caractérisations de l'utilisateur/contributeur ; et *3)* évolution des caractérisations de l'environnement de l'utilisateur/contributeur.

### 3.1 Évolution des types d'utilisateurs / contributeurs impliqués

Une première évolution concerne les types d'utilisateurs/contributeurs impliqués dans la construction collaborative des ontologies.

On constate un élargissement des types d'utilisateurs/contributeurs : de l'ontologue ou ingénieur de la connaissance (celui qui structure et formalise la connaissance) à toute partie prenante (*stakeholder*), c'est-à-dire à *« quiconque est activement impliqué dans le développement d'une ontologie et dans son usage, et quiconque dont les intérêts peuvent être*

*affectés par le développement de cette ontologie » (*di Maio, 2011) (voir Table 2a). Notons cependant que, pour certains auteurs, les ontologues ou ingénieurs de la connaissance sont exclus des parties prenantes : les parties prenantes incluent les utilisateurs pas ou peu experts en ingénierie ontologique ; d'où la remarque que ces parties prenantes ont acquis désormais un « statut de première classe » dans le processus de construction des ontologies (cf. Shosha et al., 2015).

On constate également un passage de l'individu ou du groupe restreint au collectif ouvert : de l'ontologue ou ingénieur de la connaissance ou de l'équipe d'ontologues à la communauté, voire à la foule (*crowd*) (voir Table 2b).

*TABLE 2 – Types d'utilisateurs des systèmes de construction collaborative d'ontologies (liste non exhaustive) : a) Individus ; b) Collectifs*

| **a) INDIVIDUS** | |
|---|---|
| **Noms** | **Variantes ou Instances** |
| Ontologist<br>*Ontologue* | Ontology expert \| \| Ontology engineering expert \| Ontology engineer \| Ontology designer \| Ontology builder \| Ontology developer \| Ontology author \| … |
| Knowledge Engineer<br>*Ingénieur de la connaissance* | Knowledge expert \| …<br>Ontology engineer \| Linguist) \| …<br>Cogniticien |
| Domain Expert<br>*Expert du domaine* | Expert \| Instructional design expert \| External expert \| Subject matter expert \|Alpha subject matter expert \| Domain specialist \| Domain connoisseur \| … |
| Developer | Knowledge-based system developer |
| Researcher | Knowledge-engineering researcher |
| User<br>*Utilisateur* | End User \| Targeted end-user \| Potential end-user \| Ontology user \| Application user \| Web user \| Non expert builder …<br>Direct user \|Indirect user<br>Professional \| Practitioner \| Researcher \|Social Scientist \| … |
| Stakeholder<br>*Partie prenante* | Interested Party \| Ontology stakeholders \| Domain knowledge stakeholder…<br>Users \|Sponsors \| Investors \| Technology providers \| Industry associations \| Standardization bodies \| Other people and roles |

| **b) COLLECTIFS** | |
|---|---|
| **Noms** | **Variantes ou Instances** |
| Dyad<br>*Dyade* | Dyads having different cognitive styles |
| Group<br>*Groupe* | Single research group \| Group of contributors \| Group of people \| Group of non experts \|<br>Group of users with diverse levels of ontology expertise and training… \|Beginner user g roupe \| Expert user group \| …<br>Stakeholder group \|Stakeholder subgroup \| Ontology- co-creation stakeholder group \| …<br>Group of editors \|Editor group \| Guest group \| |
| Board<br>*Comité* | Expert board \| Board of ontology stakeholders \| … |
| Team<br>*Equipe* | Design team \|Team of ontology engineers and domain experts \| … |
| Project<br>*Projet* | Project participants \| Participants to an ontology engineering project \| … |
| Organization<br>*Organisation* | International organization \| … |

| | |
|---|---|
| Community<br>*Communauté* | Community members \|Community of learners \|Large user community \| Community of domain experts \| Intra-community \| Co-evolving communities… |
| Consortium<br>*Consortium* | Consortium of ontology and software developers \| … |
| Network<br>*Réseau* | Réseau interdisciplinaire \|… |
| Population<br>*Population* | Large population of non experts \| … |
| Crowd<br>*Foule*<br>*[Grand Public]* | Crowd worker \| … |

## 3.2 Évolution des caractérisations de l'utilisateur/contributeur

Une autre catégorie d'évolutions concerne les caractérisations de l'utilisateur/contributeur. Ces caractérisations deviennent plus complètes et plus détaillées. Ces évolutions montrent le besoin de mieux identifier et comprendre les types possibles d'utilisateurs/contributeurs afin de mieux adapter les systèmes à ces derniers. L'évolution des caractérisations s'observe du point de vue du contenu et du format de ces dernières.

### 3.2.1 Contenu des caractérisations de l'utilisateur

Du point de vue du contenu des caractérisations, on constate ainsi :

- *une description de plus en plus fondée sur des observations de contributeurs réels* ; par analogie avec la distinction chère aux ergonomes entre tâches prescrites et tâches effectives – ces dernières étant fondées sur des observations –, on pourrait introduire la distinction entre *modèle prescrit* et *modèle effectif* ; d'ailleurs, Van Laere et al. (2014) montrent que les types et rôles prescrits de parties prenantes ne suffisent pas à caractériser les contributeurs d'un projet ontologique et qu'il est nécessaire de les établir grâce par exemple une méthode de profilage à partir d'une analyse des interactions réelles entre ces parties prenantes et entre chaque partie prenante et le système ;
- *une description de plus en plus fine des rôles de chaque contributeur* (rôle explicite, rôle implicite ; rôle reconnu, rôle non (encore) reconnu ; etc.). Sont apparus par exemple les rôles de : *Super ontologist* | *Tacit ontologist* ; *Ontology author* ; *Ontology committer* (*a proxy for author*) | *Unique committer* ; *Ontology manager* ; *Ontology editor*; *Ontology submitter* ; *Ontology curator* | *Central curator* | *Single curator* | *Alpha curator* | *Content curator*; *Validator*; *Scribe* | *Single scribe*; *Modeler* | *Individual modeler*; *Facilitator* | *Moderator* ; *Panelist* ; *Authoritative party* ;
- *une description plus poussée des motivations, buts, engagements, connaissances, compétences, préférences, exigences, attitudes, émotions… des utilisateurs/contributeurs et de leur impact sur les systèmes et sur les méthodes*. Par exemple, Mezghani et al. (2016) ont observé que les utilisateurs ayant une formation en génie logiciel préféraient une approche asynchrone de l'ingénierie ontologique collaborative alors que les utilisateurs ayant une formation en science informatique préféraient une approche synchrone ;
- *une description plus complète des tâches, modes opératoires des différentes catégories d'utilisateurs/contributeurs* ; des activités jusque-là considérées comme secondaires ou les activités auxquelles on n'avait pas pensé sont mises en valeur (ex. : les retours des utilisateurs des ontologies). En rapport avec cette prise en compte de ce type d'activités est apparue la notion d'« ontologues tacites » (*tacit ontologists*), c'est-à-dire ceux qui contribuent par des mails ou par des annotations (Malone & Stevens,

2013). On se dirige ainsi vers une reconnaissance de toutes les activités contributrices ;
- *une description plus poussée des modes de collaboration et d'interaction entre contributeurs* (ces derniers participant au même projet).

### 3.2.2 Format des caractérisations de l'utilisateur

Du point de vue du format des caractérisations, on observe entre autres : un *passage d'une représentation abstraite à une représentation plus concrète de l'utilisateur/contributeur* (ex. : représentation sous forme de Personas – voir section 4.1.1) ; une *description plus fréquente de scénarios impliquant les utilisateurs/contributeurs* (voir section 4.1.1) ; une *inclusion plus fréquente d'une représentation de l'utilisateur/contributeur dans les schémas d'architecture* des systèmes de construction collaborative d'ontologies (cf. de Sainte Marie et al., 2011).

## 3.3 Évolution des caractérisations de l'environnement de l'utilisateur/contributeur

Une troisième catégorie d'évolutions concerne l'environnement physique et social (ou cadre [*setting*] ou contexte) de l'utilisateur/contributeur, cet environnement exerçant une influence sur l'activité des utilisateurs/contributeurs. De manière générale, on observe une description de l'utilisateur/contributeur incluant de plus en plus de caractéristiques de cet environnement. On constate par exemple :

- une *description incluant de plus en plus des caractéristiques sociales* : mode de répartition du travail ou mode d'attribution d'une contribution (en relation avec les rôles des contributeurs ; cf. Li et al., 2005), mode de communication (Toppano, 2010), particularités culturelles (en particulier lorsque la collaboration est interculturelle ; cf. Anticoli & Toppano, 2011a et b)… ;
- une *description de la contribution de l'utilisateur/contributeur en rapport avec la tâche principale ou courante de ce dernier* (la tâche principale ou courante motivant la contribution à l'ontologie ; voir par exemple l'association entre *tagging* (folksonomie) et construction de l'ontologie ; Huyinh-Kim-Bang et al., 2008 ; Limpens et al., 2008).

# 4 Évolutions parallèles

Nous décrivons maintenant des évolutions qui ont eu lieu parallèlement à l'évolution du modèle de l'utilisateur/contributeur : *1)* l'évolution des méthodes de conception des systèmes (en particulier des méthodes d'analyse et de modélisation des utilisateurs) ; *2)* l'évolution des systèmes de construction collaborative d'ontologies ; et *3)* l'évolution des méthodes de construction collaborative d'ontologies. L'objectif de cette section est de montrer les conséquences ou les causes de l'évolution du modèle de l'utilisateur/contributeur.

## 4.1 Évolution des méthodes de conception des systèmes

Les méthodes de conception des systèmes, en particulier les méthodes d'analyse et de modélisation des utilisateurs, se sont davantage centrées sur l'utilisateur. Les concepteurs se sont ainsi beaucoup inspirés des méthodes utilisées dans les communautés IHM et CSCW (*Computer-Supported Cooperative Work*). L'objectif était de mieux comprendre les utilisateurs/contributeurs pour mieux adapter les systèmes à ces derniers et guider les concepteurs. On observe un développement des approches empiriques et théoriques.

### 4.1.1 Approches empiriques

On constate ainsi par exemple :

- un fort développement des ***études empiriques de l'activité de construction collaborative des ontologies*** dans des projets réels avec des systèmes existants : analyse et modélisation des *patterns (logs) d'usage* dans des projets de taille et d'envergure différentes (Walk et al., 2014) ; étude des *patterns collaboratifs* dans des grands projets de développement d'ontologies (Falconer et al., 2011) ; étude de la *dynamique de collaboration* lors de la construction d'ontologies avec des environnements différents – WebProtégé et MoKi (Rospocher et al., 2014) ; étude des *processus d'ingénierie collaborative distribuée* et des capacités correspondantes de l'outil Collaborative Protégé (Schober et al., 2009) ; étude de plusieurs projets de développement d'ontologies sous l'angle des *processus et* des *coûts* (Simperl & Tempich, 2006) ; étude des problèmes rencontrés dans les différentes tâches de construction d'ontologies (Vigo et al., 2014) : étude de la *construction distribuée d'ontologies comme pratique professionnelle* (Randall et al., 2011) ; étude exploratoire de l'*ingénierie ontologique pour la documentation en architecture* logicielle (de Graaf et al., 2014) ; étude de l'*impact sur la qualité de l'ontologie de l'implication d'experts du domaine dans l'annotation sémantique* de leurs articles, par comparaison avec l'annotation par des ingénieurs de la connaissance (Tatarintseva & Ermolayev, 2013) ; étude des *activités de co-création d'une ontologie* par différentes parties prenantes (Bleumers et al., 2011) ; étude du *rôle de la répartition géographique*, etc., sur la construction d'ontologies (Pinto et al., 2009) ; étude de la *motivation des utilisateurs/contributeurs* – c'est ainsi que les experts du domaine (auteurs d'articles) ne sont pas motivés pour s'engager à raffiner l'ontologie car ils ne sont pas impliqués dans le processus d'annotation, lequel est pris en charge par l'ontologue (Tatarintseva & Ermolayev, 2013) ; étude des *facteurs déterminant l'implication des différents contributeurs* dans la construction collaborative des ontologies (Randall et al., 2011), des facteurs tels que le moment auquel participer ou le lien avec ses propres objectifs ou attentes (cf. l'attitude de « satisfaction suffisante » ou *satisficing attitude* : ce que font les contributeurs est « suffisamment bon » pour atteindre leurs objectifs) ;
- l'utilisation de la *technique des défis*; voir par exemple le *Collaborative Knowledge Construction (CKC) Challenge* (Noy et al., 2008) organisé dans le but de faire tester différents outils de construction collaborative de connaissances à différents utilisateurs et d'apprendre ce que ces utilisateurs attendent de ce type d'outils ;
- l'utilisation de *techniques d'analyse et de définition de profils utilisateurs* en ingénierie ontologique collaborative (Van Laere et al., 2014) ;
- l'utilisation de *techniques d'analyse et de modélisation de l'utilisateur* classiques en IHM, comme la *technique des Personas* d'Alan Cooper (1999) ; les personas sont des archetypes *concrets* d'utilisateurs élaborés à partir de données d'entretiens et/ou d'observations d'utilisateurs potentiels du dispositif à concevoir ou à évaluer ; cf. par exemple l'utilisation de cette technique par de Bonis et al. (2011 ; de Sainte Marie et al., 2011) pour évaluer l'utilisabilité du composant de création d'ontologies de la plate-forme OntoRule par des chefs d'entreprise, des analystes d'affaire et des développeurs ;
- l'utilisation des *techniques d'analyse des parties prenantes* pour rendre compte de la collaboration entre multiples parties prenantes (Kozaki et al., 2011). NB : la technique des Personas peut être considéré comme une technique d'analyse de parties prenantes dans la mesure où elle fait la distinction entre personas primaires, personas secondaires, personas tertiaires et *ante-personas* ;
- l'utilisation de *techniques de définition de rôles* (cf. Li et al., 2005 ; Walk et al., 2014) ;
- l'utilisation des *techniques de scénarios d'usage ou de tâches* (cf. Giboin et al., 2002 ; Palma et al., 2011). On notera que la technique des Personas suppose l'élaboration de scénarios mettant en scène les personas utilisant le système en cours de conception ;

- l'utilisation d'*indicateurs centrés utilisateurs* tels que le degré de participation ou le degré d'agrément/accord avec la représentation du domaine (*degree of community grounding*) (Siorpaes & Hepp, 2007) ;
- l'élaboration d'*outils de visualisation des processus de construction collaborative de différents contributeurs*, tels que l'outil Pragmatix (Walk et al., 2013 ; Walk et al., 2014 ; Walk et al., 2016).

#### 4.1.2 Approches théoriques

A côté de ces approches empiriques, se développent également des approches plus théoriques. Ces approches font appel à des théories des sciences humaines et sociales pour modéliser les utilisateurs/contributeurs. Par exemple :

- *Communication entre concepteurs et utilisateurs des ontologies :* Toppano (2010) utilise un modèle basé sur la communication pour décrire la conception et l'utilisation ou la réutilisation des ontologies ; dans ce modèle les ontologies sont considérées comme des objets sémiotiques et l'accent est mis sur la relation entre l'interprétation de ces objets par les concepteurs des ontologies et leur interprétation par les utilisateurs.
- *Influence des contributeurs :* Considérant l'ontologie comme un objet social, Aimé et Charlet (2012, 2016) font appel à la psychologie sociale pour déterminer les influences des différents contributeurs (ontologue, expert du domaine) ; ils font référence en particulier à la psychologie socio-sémiotique de Chabrol (1984) et à la conceptualisation du Web socio-sémantique de Zacklad (2005).
- *Traits de personnalité des contributeurs :* Mezghani et al. (2016) se basent sur des travaux de psychologie montrant que les comportements de partage de connaissance entre individus sont influencés par les traits de personnalité de ces individus. Pour déterminer les personnalités des utilisateurs/contributeurs et éviter les problèmes de communication entre ces derniers, ils utilisent le modèle des *Big Five*, qui décrit les cinq traits centraux de la personnalité (Ouverture, Conscienciosité, Extraversion, Agréabilité, Névrosisme).

### 4.1 Évolution des systèmes de construction collaborative d'ontologies

Les systèmes (leurs fonctionnalités, leurs interfaces utilisateurs, etc.) ont évolué vers une plus grande adaptation aux différents types d'utilisateurs/contributeurs possibles, à leurs caractéristiques et à leurs activités, en particulier collaboratives. On observe entre autres :

- l'*élargissement du périmètre fonctionnel des systèmes*. Au départ on parlait d'*éditeurs* ; on parle davantage maintenant d'*environnements* ou de *plates-formes*. Cet élargissement peut se traduire par une connexion à d'autres outils que ceux contenus dans l'environnement proprement dit de construction collaborative d'ontologies ;
- le *développement de fonctionnalités pour la réalisation de tâches amont et aval de construction de l'ontologie*, tâches auxquelles peuvent contribuer les parties prenantes autres que les ontologues. Par exemple : *a)* préconisation d'une phase de pré-conceptualisation dans le processus d'ingénierie ontologique collaborative (Braun et al., 2007) ; *b)* aide à la modélisation, par les experts du domaine, des différences de points de vue sur le sens des termes (Towntology ; Keita et al., 2006) ; *c)* aide à l'extraction contextuelle des termes candidats (ECCO ; Giboin & Durville, 2007) ;
- l'i*mplémentation de* plug-ins *collaboratifs pour des environnements non collaboratifs*, par exemple pour Protégé (d'où Collaborative Protégé ; Tudorache & Noy, 2007), pour WebOnto (d'où Hozo ; Kozaki et al., 2007) ou pour Terminae (*plug-in* en projet ; Ressad-Boudghaghen et al., 2013) ;
- la *connexion à des outils de représentation moins formelle des connaissances* (langage semi-formel de graphes, langue naturelle..) ou le *développement de*

*fonctionnalités de visualisation des ontologies* : cf. CMapTools (Cañas et al., 2004), VocBench (Stallato et al., 2015), Bio-Mixer (Fu et al., ), Hozo (Kozaki et al., 2007), WebOnto et Tadzebao (Domingue, 1998), ces deux derniers outils permettant de mettre en œuvre une « ingénierie ontologique graphique », par exemple d'exprimer des arguments à l'aide de textes, d'images Gif et même de croquis dessinés à la main ;

- l'*utilisation de traducteurs pour passer d'un langage à un autre* : comme les traducteurs vers Prolog, CLIPS et LOOM dans le cas d'Ontolingua (Farquhar et al., 1997) ou les traducteurs vers RDF(S) et OWL dans le cas de WebODE (Corcho et al., 2002) ;
- le *développement de fonctionnalités de communication entre contributeurs* ; pour cela on se base beaucoup sur les wikis (cf.. Siorpaes & Hepp, 2007, MyOntology ; Mezghani et al., 2016 ; Semantic Media Wiki) ;
- l'*intégration d'une fonctionnalité d'argumentation*, pour aider par exemple aux délibérations entre contributeurs, comme dans Compendium (Buckingham Shum et al. 2002) ou dans WebOnto et Tadzebao (Domingue, 1998) ;
- l'*intégration d'une fonctionnalité de recommandation*, pour recommander par exemple des concepts aux experts du domaine (Walk et al., 2012).

### 4.2 Évolution des méthodes de construction collaborative d'ontologies

Les méthodes de construction collaborative d'ontologies ont également évolué vers une plus grande adaptation aux différents types d'utilisateurs/contributeurs possibles, à leurs caractéristiques et à leurs activités, en particulier collaboratives. Là encore, les « méthodologues » ont cherché à centrer leurs méthodes sur les utilisateurs/collaborateurs réels de ces méthodes. Deux de ces « méthodologues » ont d'ailleurs intitulé leur méthode « méthodologie d'ingénierie ontologie centrée sur l'humain » – *Human-Centered Ontology Engineering. Methodology*, HCOME (Kotis & Vouros, 2006).

Plusieurs directions d'évolution « centrées utilisateurs/contributeurs » apparaissent, parmi lesquelles :

- *simplifier la démarche de construction des ontologies*, comme dans le *« Just enough » ontology engineering* de di Maio (2011), qui permet aux utilisateurs ayant peu ou pas d'expertise en ingénierie ontologique de participer à la construction d'ontologies légères ; ou comme dans UPON Lite, la méthodologie d'ingénierie ontologique rapide (De Nicola & Missikof, 2016). Voir aussi Siorpaes & Hepp (2007) ;
- *impliquer l'ensemble des utilisateurs/contributeurs dans toutes les étapes du cycle de construction des ontologies*, comme dans HCOME (Kotis & Vouros, 2006) ou DILIGENT (Pinto et al., 2009). On notera que pour certains « méthodologues » comme Ongenae et al. (2013), la méthode HCOME est une méthode « extrémiste » car elle privilégie le rôle des experts du domaine dans les différentes phases de conception des ontologies ; Ongenae et al. proposent en conséquence une méthode occupant une position médiane entre HCOME et l'autre type de méthode extrémiste qui, elle, privilégie le rôle de l'ingénieur de la connaissance (les experts du domaine n'étant impliqués que dans la phase de spécification de l'ontologie). Voir également l'approche dite d'« ingénierie ontologique participative » (Giboin et al., 2008 ; Ongenae et al., 2011, 2013) ;
- *intégrer toutes les tâches contributives* comme dans la méthode de développement collaboratif inter-organisationnel d'ontologies de Palma et al. (2011), laquelle applique une approche holistique ;
- *s'appuyer sur le* tagging *collaboratif ou l'élaboration de folksonomies*, activités réalisées « naturellement » voire « sauvagement » par les utilisateurs/contributeurs autres que les ontologues (cf. par exemple : Halpin et al., 2007 ; Zacharias &

Braun, 2007 ; Huynh-Kim-Bang & Dané, 2008 ; Limpens et al., 2008 ; Gandon & Giboin, 2008) ;

- *développer les tâches de* crowdsourcing *ontologique* (cf. Mortensen, Musen & Noy, 2013) ;
- *s'appuyer sur une meilleure connaissance des caractéristiques cognitives et/ou affectives des utilisateurs/contributeurs*, comme dans la méthode de Gavrilova & Leschcheva (2014), qui s'appuie sur les styles cognitifs des utilisateurs/contributeurs (Dépendance vs Indépendance à l'égard du Champ, Impulsivité vs Réflexivité, Catégorisation en Largeur vs en Profondeur) ou comme dans la méthodologie de Mezghani et al. (2016) qui s'appuie sur les traits de personnalité (Ouverture, Conscienciosité, Extraversion, Agréabilité, Névrosisme) ;
- *s'appuyer sur une meilleure connaissance des disciplines des utilisateurs/contributeurs* (cf. Kotis & Vouros, 2006 ; Bourcier et al., 2006 ; Ressad-Bouidghaghen et al., 2013) ;
- *exploiter de manière plus systématique la méthode des scénarios*, par exemple dans la phase d'évaluation des ontologies, cette méthode étant utilisée en complément de la technique des « scénarios motivants » de la méthode TOVE de Grüninger et ses collègues (cf. Giboin et al., 2002). Voir également Mezghani et al. (2016) ;
- *favoriser l'interactivité* comme dans la méthode interactive LOVMI de validation structurelle et sémantique des ontologies de Richard et al. (2015) ;
- *tracer la logique de conception des ontologies* de façon à améliorer l'intercompréhension des décisions de conception (Dellschaft et al., 2008).

## 5 Perspectives d'évolution envisagées par les concepteurs eux-mêmes

Nous venons de voir comment a évolué le modèle de l'utilisateur/contributeur de systèmes de construction collaborative d'ontologies et comment ont évolué en parallèle les méthodes de conception de ces systèmes, les systèmes eux-mêmes ainsi que les méthodes de construction collaborative d'ontologies. Quelles sont maintenant les perspectives d'évolutions futures qu'envisagent les concepteurs en rapport avec ces différents axes d'évolution ? Certaines de ces perspectives ont été réalisées et présentées dans les sections précédentes. On décrit ici quelques perspectives non encore réalisées à ce jour, ou pas entièrement.

**Perspectives d'évolution du modèle de l'utilisateur/contributeur.–** *a) Evolution des types d'utilisateurs/contributeurs impliqués :* préciser mieux les types de collectifs ; *b) Évolution des caractérisations de l'utilisateur/contributeur :* Enrichir les descriptions de tâches/stratégies/… des différents contributeurs; de leurs rôles ; des intérêts ; de leurs exigences sur les interfaces (Tudorache et al., 2008, 2013) ; mieux définir la notion de communauté : sa dénomination, sa composition, son hétérogénéité, les relations entre ses membres (Randall et al., 2011) ; *c) Evolution des caractérisations de l'environnement de l'utilisateur/contributeur :* définir les cadres réels d'utilisation des systèmes collaboratifs (Tudorache et al., 2008, 2013) ; décrire les mauvais usages et les actes de vandalisme afin de les prévenir (Mainz et al., 2008).

**Perspectives d'évolutions parallèles.–** *a) Évolution des méthodes de conception des systèmes* : découvrir davantage les utilisateurs/contributeurs ; découvrir de nouveaux résultats surprenants sur ces utilisateurs/contributeurs ; analyser la dynamique du développement collaboratif d'ontologies ; développer la connaissance des rôles ; évaluer le système dans d'autres cadres réels que ceux dans lesquels le système a déjà été testé (Tudorache et al., 2008, 2013) ; *b) Évolution des systèmes de construction collaborative des ontologies :* ajouter des mécanismes supplémentaires de collaboration ; accorder aux utilisateurs/contributeurs des privilèges différents à différents niveaux de granularité en fonction des rôles (Tudorache et al., 2008, 2013) ; c) *Évolution des méthodes de construction collaborative d'ontologies :* supporter des workflows différents selon les différents types d'utilisateurs/contributeurs (Tudorache et al., 2008, 2013).

## 6 Conclusion

Nous venons de rendre compte d'une étude en cours sur l'évolution du modèle de l'utilisateur de systèmes de construction collaborative d'ontologies et sur les évolutions parallèles des méthodes de conception des systèmes de construction collaborative d'ontologies, des systèmes eux-mêmes et des méthodes de construction collaborative des ontologies. Cette étude se poursuit actuellement. Elle mériterait d'ailleurs d'être poursuivie de manière collaborative avec les membres de la communauté IC.

Quoi qu'il en soit, notre intention en présentant cette étude en cours était de montrer qu'une meilleure connaissance des utilisateurs/contributeurs – reflétée par un modèle de l'utilisateur/contributeur plus réaliste – est importante pour concevoir des systèmes adaptés. En 2013 Tudorache et al. écrivaient : *« À mesure que nous en apprenons plus sur la manière dont les experts du domaine construisent des ontologies dans un environnement distribué, nous pouvons ajuster les outils pour améliorer la collaboration. »* Avec la présente étude, on pourrait écrire : À mesure que nous en apprenons plus sur la manière dont tous les types de contributeurs construisent des ontologies dans un environnement distribué, nous pouvons ajuster les outils pour améliorer la collaboration entre ces contributeurs.

Cette étude a fourni par ailleurs des indications sur : a) les méthodes à utiliser pour acquérir cette meilleure connaissance de l'utilisateur ; b) les ajustements des méthodes de conception des systèmes de construction collaborative d'ontologies, des systèmes eux-mêmes et des méthodes de construction collaborative des ontologies, auxquels conduit cette connaissance.

## Références


AIME X, CHARLET J. IC : Ingénierie des Connaissances ou Ingénierie du Conformisme ? In: Troncy R, editor. *Actes des 24es journées francophones d'Ingénierie des Connaissances (IC'2013)*. Lille, France ; 2013.

AIMÉ X, CHARLET J. (2016). Social Psychology Insights into Ontology Engineering. *Future Generation Computer Systems* 54, p. 348–351.

ANTICOLI, L., TOPPANO, E. (2011a). The role of culture in collaborative ontology design. ISWSA 2011: 4.

ANTICOLI, L., TOPPANO, E. (2011b). How Culture May Influence Ontology Co-Design: A Qualitative Study. *IJITWE 6(2)*, p. 1-17.

AUER S., DIETZOLD S., & RIECHERT T. (2006) OntoWiki – A Tool for Social, Semantic Collaboration. In Cruz I. et al. (eds) *The Semantic Web - ISWC 2006. ISWC 2006*. Lecture Notes in Computer Science, vol 4273, p. 736-749, Springer, Berlin, Heidelberg.

BAO, J. & HONAVAR V.(2004). Collaborative ontology building with wiki@nt - a multiagent based ontology building environment. In *Proc. of 3rd International Workshop on Evaluation of Ontology-based Tools*, located at ISWC 2004, 8th November 2004, Hiroshima, Japan, p. 37–46.

BAO, J., HU, Z., CARAGEA, D., REECY, J., & HONAVAR, V. (2006). A tool for collaborative construction of large biological ontologies. In: Bressan, S., Küng, J., Wagner, R. (eds.) *DEXA 2006*. LNCS, vol. 4080, pp. 191–195. Springer, Heidelberg.

BAO, J., HU, Z., D., REECY, J., HONAVAR, V.G. (2006). A Proposal for Collaborative Ontology Editor for Animal Trait Ontology.

BLEUMERS, L., JACOBS, A., ONGENAE, F.,ACKAERT, A., SULMON, N., VERSTRAETE, M., VAN GILS, M., & DE ZUTTER, S.. (2011). Towards Ontology Co-creation in Institutionalized Care Settings. *5th International Conference on Pervasive Computing Technologies for Healthcare and Workshops, PervasiveHealth 2011,* p. 559-562.

BOURCIER, D., DULONG DE ROSNAY, M., & LEGRAND, J. (2006). Susciter la construction interdisciplinaire d'ontologies juridiques : bilan d'une expérience. In M. Harzallah, J. Charlet et N. Aussenac-Gilles. *Semaine de la Connaissance, journée Ontologies et textes juridiques*, Juin 2006, Nantes, France. 3, p. 50-59.

BRAUN, S., SCHMIDT, A., WALTER, A., NAGYPAL, G., & ZACHARIAS, V. 2007. Ontology maturing: a collaborative Web 2.0 approach to ontology engineering. In *Proceedings of the Workshop on*

*Social and Collaborative Construction of Structured Knowledge* at the 16th International World Wide Web Conference (WWW 2007), Edinburgh.

BUCKINGHAM SHUM, S., MOTTA, E., & DOMINGUE, J. (2002) Augmenting Design Deliberation with Compendium: The Case of Collaborative Ontology Design. *Workshop on Facilitating Hypertext-Augmented Collaborative Modelling* ACM Hypertext Conference, Maryland, June 11th-12th, 2002.

CAÑAS, A. J., HILL, G., CARFF, R., SURI, N., LOTT, J., ESKRIDGE, T., et al. (2004). CmapTools: A knowledge modeling and sharing environment. In A. J. Cañas, J. D. Novak & F. M. González (Eds.), *Concept maps: Theory, methodology, technology. Proceedings of the first international conference on concept mapping* (Vol. I, pp. 125-133). Pamplona, Spain: Universidad Pública de Navarra.

CHABROL C. (1984). Psycho-socio-sémiotique : Définitions et propositions. *Langage et société 28(1)*, p. 53-71

CHEN, Y., ZHANG, S., PENG, X., & ZHAO, W. (2008). A Collaborative Ontology Construction Tool with Conflicts Detection. In *2008 Fourth International Conference on Semantics, Knowledge and Grid*, Beijing, 2008, p. 12-19.

COOPER, A. (1999). *The Inmates Are Running the Asylum: Why High Tech Products Drive Us Crazy and How to Restore the Sanity*. Macmillan Publishing Co., Inc. Indianapolis, IN, USA.

CORBY, O. & DURVILLE, P. (2009). ECCO (Editeur Collaboratif et Contextuel d'Ontologies). *Journée "Composition logicielle"*, Sophia Antipolis, 17 avril 2009.

CORCHO, O., FERNÁNDEZ-LÓPEZ, M., GÓMEZ-PÉREZ, A., & VICENTE, O. (2002): WebODE: An Integrated Workbench for Ontology Representation, Reasoning, and Exchange. *EKAW 2002*, p. 138-153.

CORRENDO, G. & ALANI, H. (2007). Survey of Tools for Collaborative Knowledge Construction and Sharing. *Web Intelligence/IAT Workshops 2007*, p. 7-10.

DABROWSKI, M., KRUK, S.R., PIOTROWSKI, P., SZCZECKI, P. & WOZNIAK, M. (2007). Collaborative Ontology Development with MarcOnt Portal, *ESWC 2007*.

DE BONIS, S., BELLINO, C., & EL GHALI, A. (2011) *Final Usability Report: Evaluation and Conclusions*. OntoRule Project Deliverable D2.5, December 2011, 79 p.

DE GRAAF, K.A. LIANG, P., TANG, A., VAN HAGE, W.R., & VAN VLIET, H. (2014). An exploratory study on ontology engineering for software architecture documentation. *Computers in Industry 65(7)*, p. 1053-1064 .

DELLSCHAFT, K., ENGELBRECHT, H., BARRETO, J.M, RUTENBECK, S., & STAAB, S.(2008). Cicero: Tracking Design Rationale in Collaborative Ontology Engineering. ESWC 2008. p. 782-786.

DE NICOLA, A. & MISSIKOFF, M.: (2016). A lightweight methodology for rapid ontology engineering. *Commun. ACM 59(3)*, p. 79-86.

DE SAINTE MARIE C., IGLESIAS ESCUDERO M., & ROSINA P. (2011) The ONTORULE Project : Where Ontology Meets Business Rules. In: Rudolph S., Gutierrez C. (Eds.) *Web Reasoning and Rule Systems* (p. 24-29), RR 2011. Lecture Notes in Computer Science, vol 6902. Springer, Berlin, Heidelberg.

DIAMANTINI, C., GENGA, L., POTENA, D., & STORTI, E. (2014). Collaborative Building of an Ontology of Key Performance Indicators. *OTM Conferences 2014*, p. 148-165

DI MAIO, P. (2011). 'Just enough' ontology engineering. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics (WIMS '11)*, Article No. 8.

DOMINGUE, J. (1998). Tadzebao and WebOnto: Discussing, Browsing, and Editing Ontologies on the Web. In *Eleventh Workshop on Knowledge Acquisition,Modeling and Management, EKAW 98* (18-23 April 1998, Banff, Alberta, Canada)

ERDMANN M., WATERFELD W. (2012) Overview of the NeOn Toolkit. In: Suárez-Figueroa M., Gómez-Pérez A., Motta E., Gangemi A. (eds) Ontology Engineering in a Networked World. Springer, Berlin, Heidelberg.

FALCONER, S.M., TUDORACHE, T., & NOY, N.F. (2011). An analysis of collaborative patterns in large-scale ontology development projects. *K-CAP 2011*, p. 25-32

FARQUHAR, A., FIKES, R., & RICE, J. (1997). The ontolingua server: A tool for collaborative ontology construction. *International journal of human-computer studies 46 (6)*, p. 707-727.

FU, B., GRAMMEL, L., & STOREY, M-A.D. (2012). BioMixer: A Web-based Collaborative Ontology Visualization Tool. *ICBO 2012*.

GANDON, F. & GIBOIN, A. (2008). Vers des ontologies à l'état sauvage. In *Atelier IC 2.0* (associé aux 19èmes Journées Francophones d'Ingénierie des Connaissances, IC2008), Nancy.

GAVRILOVA, T. A., & LESHCHEVA, I. A. (2014). Collective Ontologies Design and Development. In Proceedings of 2014 *Eighth International Conference on Complex, Intelligent and Software Intensive Systems (CISIS– 2014)*, Birmingham, United Kingdom, IEEE Computer Society, Conference Publishing Services (CPS) (p. 564-569).

GENNARI, J.H. , MUSEN, M.A. , FERGERSON, RW., GROSSO, W.E., CRUBÉZY, M., ERIKSSON, H., NOY, N.F., & TU, S.W. (2003) The Evolution of Protégé: An environment for knowledge-based systems development. *International Journal of Human-Computer Studies 58 (1)*, p. 89-123.

GHIDINI, Ch., ROSPOCHER, M., & SERAFINI, L. (2010). MoKi: a Wiki-based Conceptual Modeling Tool. EKAW (Posters and Demos) 2010

GIBOIN, A., DURVILLE, P. (2007). [ECCO:] Editeur collaboratif et contextuel d'ontologie. In INRIA Sophia Antipolis, EADS, LISI (Eds.), Outils et Services de gestion des ontologies, Section 2 (p. 9-59), Rapport de projet ANR-RNTL « e-WoK_Hub ».

GIBOIN, A., DURVILLE, P., & GANDON, F. (2008). Ingénierie ontologique participative : essai de mise en œuvre avec l'éditeur collaboratif ECCO. In *Atelier IC 2.0* (associé aux 19èmes Journées Francophones d'Ingénierie des Connaissances, IC2008), Nancy.

GIBOIN, A., GANDON, F., CORBY, O., DIENG, R. (2002). User Assessment of Ontology-based Tools: A Step Towards Systemizing the Scenario Approach, In *Proceedings of EON'2002: Evaluation of Ontology-based Tools, OntoWeb-SIG3 Workshop at the 13th International Conference on Knowledge Engineering and Knowledge Management EKAW 2002*, Sigüenza (Spain), September 30, 2002, p. 63-73.

GIBOIN, A., GRATALOUP, S., MOREL, O., & DURVILLE, P. (2013). Building Ontologies for Analyzing Data Expressed in Natural Language, In M. Perrin, J.-F. Rainaud (Eds.), *Shared Earth Modeling, Knowledge Based Solutions for Building and Managing Subsurface Structural Models*, (p. 232-259), Technip Editions, Paris (France).

HALPIN, H., ROBU V., SHEPHERD H. (2007). The Complex Dynamics of Collaborative Tagging. In *WWW 2007*, ACM Press, p. 211-220.

HAYES, P., SAAVEDRA, R. & REICHHERZER, Th. (2003). A Collaborative Development Environment for Ontologies (CODE).

HAYES, P., ESKRIDGE, T.C., MEHROTRA, M., BOBROVNIKOFF, D., REICHHERZER, Th., & SAAVEDRA, R. (2005). COE: Tools for collaborative ontology development and reuse.

HUYNH-KIM-BANG, B. & DANE, E. (2008). Social bookmarking et tags structurés. *19es Journées Francophones d'Ingénierie des Connaissances (IC 2008)*, Juin 2008, Nancy, France. p.111-122.

JIMENEZ RUIZ, E., CUENCA GRAU, B., HORROCKS, I., & BERLANGA, R. (2011). Supporting Concurrent Ontology Development: Framework, Algorithms and Tool. *Data & Knowledge Engineering 70(1)*, p. 146-164.

KEITA, A., ROUSSEY, C., & LAURINI, R. (2006). Un outil d'aide à la construction d'ontologies pré-consensuelles : le projet Towntology. XXIVème Congrès INFORSID, May 2006, Hammamet, Tunisie. pp.911-926, 2006.

KELLY, C. & COLGAN, L. (1992). User Modeling and User Interface Design. In A. Monk, D. Diaper and D. Harrison (Eds.). *People and Computers VII* (p. 227-239). Cambridge : Cambridge University Press.

KOTIS K., VOUROS G.A., & ALONSO J.P. (2005) HCOME: A Tool-Supported Methodology for Engineering Living Ontologies. In Bussler C., Tannen V., Fundulaki I. (eds) *Semantic Web and Databases. SWDB 2004*. Lecture Notes in Computer Science, vol 3372. Springer, Berlin, Heidelberg

KOTIS, K. & VOUROS, G.A. (2006). Human-Centered Ontology Engineering: the HCOME Methodology. *International Journal of Knowledge and Information Systems (KAIS) 10(1)*, p. 109–131.

KOZAKI, K., SAITO, O., & MIZOGUCHI, R. (2012). A Consensus-Building Support System based on Ontology Exploration. IESD 2012, International Workshop at EKAW 20, Galway, Ireland

LI M., WANG D., DU X., & WANG S. (2005). Ontology Construction for Semantic Web: A Role-Based Collaborative Development Method. In: Zhang Y., Tanaka K., Yu J.X., Wang S., Li M. (Eds.). *Web Technologies Research and Development - APWeb 2005*. Lecture Notes in Computer Science, vol 3399. Springer, Berlin, Heidelberg.

LIMPENS, F., GANDON, F., & BUFFA, M. (2008). Rapprocher les ontologies et les folksonomies pour la gestion des connaissances partagées : un état de l'art.. *19es Journées Francophones d'Ingénierie des Connaissances (IC 2008)*, Juin 2008, Nancy, France. p.123-134.

LOSKYLL, M.; HECKMANN, D.; & KOBAYASHI, I. (2009). UbisEditor 3.0: Collaborative Ontology Development on the Web. In *Proceedings of Workshop on Web 3.0: Merging Semantic Web and Social Web 2009 (SW)²*, Turin, Italy, CEUR Workshop Proceedings.

MAINZ, I., WELLER, K. PAULSEN, I. MAINZ, D., & KOHL, J. (2008).Ontoverse: Collaborative Ontology Engineering for the Life Sciences. *Information & Praxis 59(2)*, p. 91-99.

MALONE, J. & STEVENS , R. (2013). Measuring the level of activity in community built bio-ontologies, *Journal of Biomedical Informatics 46(1)*, p. 5-14,

MANGIONE, G.R., MAZZONI, E., ORCIUOLI, F., & PIERRI, A. (2011). A Pedagogical Approach for Collaborative Ontologies Building. In T. Daradoumis et al. (Eds.). *Technology-Enhanced Systems and Tools for Collaborative Learning Scaffolding*, p. 135-166.

MEZGHANI, E., EXPÓSITO, E., & DRIRA, K. (2016).. A Collaborative Methodology for Tacit Knowledge Management: Application

MISSIKOFF, M., NAVIGLI, R., VELARDI, P.,: The Usable Ontology: An Environment for Building and Assessing a Domain Ontology. In *Proceedings of the First International Semantic Web Conference, (ISWC 2002)*, p. 39-53.

MORTENSEN JM, MUSEN MA, & NOY N.F. (2013). Crowdsourcing the Verification of Relationships in Biomedical Ontologies. *Proceedings of the AMIA Annual Symposium*.

NOY N.F., CHUGH A, & ALANI H. (2008). The CKC Challenge: Exploring Tools for Collaborative Knowledge Construction. *IEEE Intell Syst. 23(1)*, p. 64-68.

ONGENAE, F. DUYSBURGH, P., SULMON, N., VERSTRAETE, M., BLEUMERS, L., DE ZUTTER, S., VERSTICHEL, S., ACKAERT, A. JACOBS, A., & DE TURCK, F. (2013). An Ontology Co-design Method for the Co-creation of a Continuous Care Ontology, *Applied Ontology 2013*, p. 1–40

ONGENAE, F., BLEUMERS, L. SULMON, N., VERSTRAETE, M., VAN GILS, M., JACOBS, A., DE ZUTTER, S., VERHOEVE, P., ACKAERT, A., & DE TURCK, F. (2011). Participatory Design of a Continuous Care Ontology: Towards a User-Driven Ontology Engineering Methodology

PALMA, R., CORCHO, O., GÓMEZ-PÉREZ, A., & HAASE, P. (2011). A holistic approach to collaborative ontology development based on change management. *J. Web Sem. 9(3)*, p. 299-314.

PETERSEN, N., COSKUN, G., & LANGE, Ch. (2016). TurtleEditor: An Ontology-Aware Web-Editor for Collaborative Ontology Development. *ICSC 2016*: 183-186. [see also: Petersen, N., Similea, A., Lange, Ch., & Lohmann, S. (2017). TurtleEditor: A Web-Based RDF Editor to Support Distributed Ontology Development on Repository Hosting Platforms. Int. J. Semantic Computing 11(3): 311-324 (2017)].

PINTO, H.S. TEMPICH, Ch. & STAAB, S. (2009). Ontology engineering and evolution in a distributed world using DILIGENT. In *Handbook on ontologies* (p. 153-176), Springer, Berlin, Heidelberg

RANDALL, D., PROCTER, R., LIN, Y., POSCHEN, M., SHARROCK, W., & STEVENS, R. (2011). Distributed ontology building as practical work. . *International Journal of Human-Computer Studies 69*, p. 220-233.

RESSAD-BOUIDGHAGHEN, O., SZULMAN, S., ZARGAYOUNA, H. & PAUL, E. (2013). Construction collaborative d'une Ressource Termino-Ontologique (RTO) pour le droit des collectivités territoriales. *IC - 24èmes Journées francophones d'Ingénierie des Connaissances*, Jul 2013, Lille, France.

RICHARD M., AIMÉ X., KREBS M.O., & CHARLET J. (2015). LOVMI : vers une méthode interactive pour la validation d'ontologies. In: Actes des 26es journées francophones d'Ingénierie des Connaissances (IC'2015).

ROSPOCHER M., TUDORACHE T., MUSEN M.A. (2014) Investigating Collaboration Dynamics in Different Ontology Development Environments. In: Buchmann R., Kifor C.V., Yu J. (Eds.) *Knowledge Science, Engineering and Management. KSEM 2014*. Lecture Notes in Computer Science, vol 8793. Springer, Cham

SCHOBER, D., MALONE, J., & STEVENS, R. (2009). Practical experiences in concurrent, collaborative ontology building using Collaborative Protégé. *Nature Precedings* <http://hdl.handle.net/10101/npre.2009.3517.1> (2009)

SHOSHA R., DEBRUYNE C., & O'SULLIVAN D. (2015) Towards an Adaptive Tool and Method for Collaborative Ontology Mapping. In: Ciuciu I. et al. (Eds.) *On the Move to Meaningful Internet Systems: OTM 2015 Workshops*. Lecture Notes in Computer Science, vol 9416. Springer, Cham.

SIMPERL, E. & LUCZAK-RÖSCH, M. (2014). Collaborative ontology engineering: a survey. *Knowledge Eng. Review 29(1)*, p. 101-131.

SIMPERL, E. and TEMPICH, C., (2006) Ontology engineering: A reality check Meersman, R. and Tari, Z. (eds.) *The 5th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE2006)*, France. 31 Oct - 02 Nov 2006. , p. 836-854.

SIORPAES, K. & HEPP, M. (2007).myOntology: The Marriage of Ontology Engineering and Collective Intelligence, *Proceedings of the ESWC 2007 Workshop "Bridging the Gap between Semantic Web and Web 2.0"*, June 7, 2007, Innsbruck, Austria.

STELLATO, A., RAJBHANDARI, S., TURBATI, A., FIORELLI, M., CARACCIOLO, C., LORENZETTI, T., KEIZER, J., & PAZIENZA, M.T. (2015). VocBench: A Web Application for Collaborative Development of Multilingual Thesauri. *ESWC 2015*, p. 38-53

STELLATO, A., TURBATI, A., FIORELLI, M., LORENZETTI, T., COSTETCHI, E., LAABOUDI, Ch., VAN GEMERT, W., & KEIZER, J. (2017). Towards VocBench 3: Pushing Collaborative Development of Thesauri and Ontologies Further Beyond. *NKOS@TPDL 2017*; p. 39-52

SURE, Y., ERDMANN, M., ANGELE, J., STAAB, S., STUDER, R., & WENKE, D. (2002). OntoEdit: Collaborative ontology development for the semantic web. In: Horrocks, I., Hendler, J. (Eds.) *ISWC 2002*. LNCS, vol. 2342, p. 221. Springer, Heidelberg.

TATARINTSEVA, O. & ERMOLAYEV, V. (2013). Refining an Ontology by Learning Stakeholder Votes from their Texts. In *ICTERI: International Conference on ICT in Education, Research, and Industrial Applications* (p; 64-78), CEUR-WS.org.

TENNISON, J., O'HARA, K., & SHADBOLT, N. (2002). APECKS: using and evaluating a tool for ontology construction with internal and external KA support. *Int. J. Hum.-Comput. Stud. 56(4)*, p. 375-422.

TOPPANO, E. (2010). A communication-based model of ontology design and (re)use. ISWSA '10 Proceedings of the 1st International Conference on Intelligent Semantic Web-Services and Applications, Article No. 6 , Amman, Jordan — June 14 - 16, 2010.

TUDORACHE, T. & NOY, N. (2007). Collaborative Protégé. In *Workshop on Social and Collaborative Construction of Structured Knowledge (CKC 2007)* at WWW 2007, Banff, Canada, 2007.

TUDORACHE, T., NOY, N. F., TU, S., & MUSEN, M. A. (2008). Supporting Collaborative Ontology Development in Protégé. In A. Sheth et al. (Eds.): *ISWC 2008*, LNCS 5318, p. 17–32.

TUDORACHE, T., NYULAS, C., NOY, N. F., & MUSEN, M. A. (2013). WebProtégé: A Collaborative Ontology Editor and Knowledge Acquisition Tool for the Web. *Semantic Web*, *4*(1), p. 89–99.

VALO, A., HYVÖNEN, E., & KOMULAINEN, V. (2005). A Tool for Collaborative Ontology Development for the Semantic Web. In *Proceedings of International Conference on Dublin Core and Metadata Applications (DC 2005)*, Nov, 2005.

VAN LAERE S., BUYL R., & NYSSEN M. (2014) A Method for Detecting Behavior-Based User Profiles in Collaborative Ontology Engineering. In: Meersman R. et al. (Eds.) *On the Move to Meaningful Internet Systems: OTM 2014 Conferences. OTM 2014*. Lecture Notes in Computer Science, vol 8841. Springer, Berlin, Heidelberg.

VIGO, M; BAIL, S; JAY, C; & STEVENS, R. (2014). Overcoming the Pitfalls of Ontology Authoring: Strategies and Implications for Tool Design. *International Journal of Human-Computer Studies 72(12)*, p. 835-845.

WALK, S., PÖSCHKO, J., STROHMAIER, M., ANDREWS, K., TUDORACHE, T., NOY, N.F., NYULAS, C., & MUSEN, M.A. (2013). PragmatiX: An Interactive Tool for Visualizing the Creation Process Behind Collaboratively Engineered Ontologies. *Int. J. Semantic Web Inf. Syst. 9(1)*, p. 45-78.

WALK, S., SINGER, Ph., STROHMAIER, M., TUDORACHE, T., MUSEN, M.A., & NOY, N.F. (2014). Discovering Beaten Paths in Collaborative Ontology-Engineering Projects using Markov Chains, *Journal of Biomedical Informatics 51*, p. 254-271.

WALK, S., TUDORACHE, T., MUSEN, M.A. (2016). Visualizing User Editing Behavior in Collaborative Ontology-engineering Projects. *VOILA@ISWC 2016*, p. 68-79.

WOHLGENANNT, G., SABOU, M., & HANIKA, F. (2016). Crowd-based ontology engineering with the uComp Protégé plugin. *Semantic Web 7(4)*, p. 379-398.

XEXEO, G., VIVACQUA, A., DE SOUZA, J.M., BRAGA, B., D'ALMEIDA, Jr, J.N., ALMENTERO, B.K., CASTILHO, R., & MIRANDA, B. (2005). COE: A collaborative ontology editor based on a peer-to-peer framework. *Advanced Engineering Informatics 19(2)*, p.113-121.

ZACHARIAS, X. & BRAUN, S. (2007).. Soboleo – social bookmarking and lightweight engineering of ontologies. In *Proc. WWW 2007 Workshop on Social and Collaborative Construction of Structured Knowledge*, Banff, Canada, May 2007.

ZACKLAD, M. (2005). Introduction aux ontologies sémiotiques dans le Web Socio Sémantique. In M.-C. Jaulent (Ed.) Actes des 16èmes journées francophones d'Ingénierie des Connaissances, Grenoble: PUG.

# Focaliser l'extraction d'épisodes séquentiels à partir de traces par le contexte.

Béatrice Fuchs[1]

Université de Lyon, UJML3, CNRS, IAE, Laboratoire LIRIS, 69372 Lyon cedex 08, France
beatrice.fuchs@liris.cnrs.fr

**Résumé** : L'exploration des traces avec l'extraction d'épisodes séquentiels vise à caractériser des utilisateurs en fonction des séquences d'actions qu'ils ont réalisées dans un environnement numérique. Une des difficultés majeures de l'extraction de connaissances est la surabondance de résultats qui rend leur exploitation difficile par un expert humain chargé d'interpréter les résultats. Or, l'utilisation de connaissances *a priori* est souvent efficace d'une part pour limiter le volume de résultats produits et d'autre part pour focaliser l'analyse sur des résultats potentiellement intéressants. Nous proposons d'utiliser le contexte des actions de la trace pour limiter les résultats de la fouille, qui s'exprime sous la forme d'une contrainte qui permet d'obtenir des épisodes signifiants du point de vue du contexte. Des expérimentations dans le cadre applicatif du jeu sérieux Tamagocours pour l'apprentissage des règles juridiques de diffusion de documents montrent la pertinence de cette contrainte pour filtrer de nombreux épisodes sans intérêt avec un rappel et une précision intéressants.

**Mots-clés** : contraintes, découverte de connaissances, analyse de traces, contexte

## 1 Introduction

Des volumes importants de traces sont accumulés par les apprenants ou les joueurs dans leur environnement numérique d'apprentissage. Les traces numériques se présentent sous la forme de séquences d'actions contextualisées et temporellement situées décrivant des parcours d'utilisateurs dans un environnement numérique. L'analyse des ces traces qui témoignent de leur activité est importante afin de mieux comprendre leurs difficultés en vue de leur proposer une assistance adaptée en conséquence ou d'améliorer les outils disponibles. Une façon d'étudier ces traces est l'extraction de connaissances à partir de données (ECD), qui vise à extraire des connaissances à partir de données dans un processus *interactif* et *itératif* (Frawley *et al.*, 1992). Une des méthodes les plus adaptées à l'exploration des traces est l'extraction d'épisodes séquentiels qui prend en compte les dimensions événementielle et temporelle des traces et vise à mettre en évidence des séquences typiques d'actions réalisées par des utilisateurs afin de caractériser leur parcours ou leur comportement. Néanmoins plusieurs problèmes se posent lors de la mise en œuvre de l'ECD.

Le premier problème récurrent en fouille de données est que celle-ci produit le plus souvent une quantité très importante de résultats avec une forte redondance dont la plus grande partie est sans intérêt et occulte les résultats intéressants. L'introduction de contraintes pour contrôler la fouille et limiter les résultats de la fouille s'avère indispensable. L'expérience a montré que l'utilisation de mesures d'intérêt *objectives*, telles que le support ou la longueur (Béchet *et al.*, 2014) sont utiles pour contrôler efficacement la fouille, mais elles restent insuffisantes lorsqu'il y a beaucoup de redondance combinatoire (van Leeuwen, 2014). De nombreux travaux ont déjà abordé le sujet pour l'exploration des motifs ensemblistes, qui reste néanmoins ouvert pour les séquences qui sont relativement moins étudiées.

Le deuxième problème rencontré lors de l'exploration des traces est la prise en compte des dimensions représentées dans les traces. L'exploration d'épisode séquentiels restreint l'analyse à deux dimensions : le type d'action (ou type d'événement,)[1] et l'ordre des actions, éventuellement associées à une information temporelle plus précise telle que la date et

1. terme généralement utilisé en fouille de données

l'heure. Ces dimensions sont importantes et doivent être prises en compte mais le contexte dans lequel une séquence d'actions s'est déroulée est également important et celui-ci n'est pas pris en considération par la méthode de fouille d'épisode séquentiels. Par exemple, on s'intéresse non seulement au moment et à l'action réalisée mais également à l'utilisateur qui a réalisé une action et sur quel objet l'utilisateur a agi, *etc.* : le contexte exprime un lien fort entre les différentes actions en lui donnant un sens et exprime une continuité des actions pour la réalisation d'un objectif précis voulu par l'utilisateur.

Ces difficultés ont été mises en évidence lors de l'étude des traces du jeu Tamagocours, un jeu sérieux pour l'apprentissage des règles juridiques de diffusion de documents dans un cadre éducatif. Pour cela nous avons expérimenté l'introduction d'une contrainte supplémentaire qui n'était pas prise en charge par l'algorithme de fouille. Cette contrainte est opérationnelle au moment du pré-traitement et prend en compte le contexte sous la forme d'un ensemble d'attributs afin de ne sélectionner que les occurrences d'épisodes pertinents au sens de ce contexte.

La suite de l'article est organisée comme suit : la section 2 synthétise les principales contributions aux problématiques, puis la section 3 présente le contexte scientifique et le cadre applicatif des propositions de recherches. Enfin notre proposition est exposée au paragraphe 4 et est suivie par une expérimentation qui permet de d'évaluer sa pertinence. Enfin nous discutons et concluons sur ce travail avec quelques perspectives.

## 2 État de l'art : exploration de données temporelles

La fouille de données assure le traitement automatique de gros volumes de données pour trouver des *motifs* correspondant à des régularités dans les données. Après interprétation par un utilisateur, ces motifs permettent de construire un modèle d'un phénomène étudié. Les traces sont des données multidimensionnelles et caractérisées par deux dimensions importantes : le type d'action et à quelle moment l'action a eu lieu, sous forme d'une estampille temporelle. Plusieurs méthodes ont été conçues afin de prendre en compte ces deux dimensions. La fouille de séquences tient compte de la disposition ordonnée des données dans les méthodes d'exploration et dans la forme des résultats produits. La découverte de motifs séquentiels dans des bases de séquences s'appuie sur une base de transactions comportant une séquence d'itemsets chacune, et l'extraction consiste à y rechercher les sous-séquences fréquentes d'itemsets (Agrawal & Srikant, 1995). L'inconvénient de cette approche pour l'exploration des traces est qu'elle ne prend pas en compte la dimension *événementielle* comme dimension prioritaire : si elle est présente, elle est considérée à rang égal avec les autres dimensions. Par ailleurs, les traces ne se présentent pas sous la forme d'une base de séquence mais plutôt sous la forme d'une ou plusieurs séquences d'actions où chaque action est associée à un ensemble d'attributs et à une information temporelle précise, et tous les types d'action ne possèdent pas nécessairement les mêmes attributs. Une méthode alternative, la découverte d'épisodes fréquents introduite par (Mannila *et al.*, 1997) exploite des données sous la forme d'une séquence d'événements où chaque événement est associé à une information temporelle précise. Cette méthode est plus adaptée à l'exploration des traces car les actions composant la trace peuvent aisément être transposées en événements dont le type correspond au type d'action et l'estampille à l'information temporelle présente dans la trace (ou au moins l'ordre). Néanmoins les attributs caractérisant le contexte des actions de la trace ne sont pas pris en compte et des traitements complémentaires s'avèrent nécessaires en amont et/ou en aval de l'étape de fouille. Si quelques variantes ont été proposées dans la littérature (Cram, 2010), aucune d'elles n'est satisfaisante compte tenu des caractéristiques des traces.

Par ailleurs, la nature multidimensionnelle des données rajoute de la difficulté au problème

de la surabondance des résultats. Les mesures d'intérêt *objectives* à la base des stratégies de fouille permettent de réduire le temps et l'espace nécessaires à l'exploration des données, notamment grâce à leur propriétés. Elles s'appuient uniquement sur les données explorées et ne présupposent aucune connaissance supplémentaire sur les données à traiter. Parmi les mesures objectives les plus classiques on peut citer le support et la confiance qui sont connues pour être insuffisants si l'on veut extraire des informations utiles et intéressantes (Geng & Hamilton, 2007). La compacité des résultats avec la propriété de fermeture des motifs a montré sa capacité à limiter fortement les résultats mais reste toutefois toujours insuffisante. Conjointement aux mesures objectives, il est également possible d'exploiter les connaissances *a priori* que l'utilisateur possède sur les données sous la forme de mesures d'intérêt *subjectives*. Si les mesures d'intérêt subjectives s'avèrent souvent très efficaces pour réduire considérablement le volume de résultats extraits par la fouille, elles ne sont pas toujours représentables simplement. Elles peuvent être difficiles à prendre en main par l'utilisateur qui doit apprendre comment formuler ses connaissances sous une forme appropriée afin de les transposer en une ou plusieurs mesures (Geng & Hamilton, 2007). Par ailleurs elles ne sont pas toujours transposables facilement dans un autre domaine d'application ou manquent parfois de généricité si bien qu'il est délicat de les prendre en compte dans l'algorithme de fouille. On peut ajouter que, si on trouve dans la littérature de nombreux travaux qui traitent des règles d'association, il est en revanche plus rare de trouver des travaux qui s'intéressent aux épisodes séquentiels, si ce n'est pour des besoins spécifiques (Perer & Wang, 2014).

Dans le cadre de l'exploration de traces, nous proposons d'une part d'exploiter l'extraction d'épisodes séquentiels et conjointement d'utiliser les connaissances *a priori* sur les traces afin de mieux préparer les données au moment du pré-traitement. Ces connaissances s'expriment de façon très simple et générique, et permettent de segmenter une trace en plusieurs séquences avant la fouille. Nous présentons dans la section suivante le cadre de ce travail.

## 3 Extraction de connaissances à partir de traces

Nous avons développé DISKIT afin d'explorer les traces. DISKIT peut être vu comme un laboratoire d'analyse des traces destiné à être complété par d'autres méthodes en amont et/ou en aval afin d'étudier des problématiques plus précises autour des traces, en particulier (mais pas seulement) des traces d'apprenants. DISKIT met en œuvre les étapes de pré-traitement et de post-traitement du processus d'ECD et encapsule l'étape de fouille assurée par DMT4SP [2] (Nanni & Rigotti, 2007), un prototype d'extraction d'épisodes séquentiels et de règles séquentielles à un conséquent à partir d'une ou plusieurs séquences d'événements. DISKIT prend en charge plusieurs options et contraintes en complément de DMT4SP. Les données de la trace sont collectées soit sous la forme d'un fichier texte soit à partir d'un *système à base de traces* (Champin *et al.*, 2013) qui sert à la fois d'entrepôt destiné au stockage, à la manipulation et la modélisation explicite des traces, mais aussi de base de connaissances afin de mémoriser les interprétations faites de phénomènes étudiés à partir des traces. Nous proposons d'expérimenter avec DISKIT la prise en compte du contexte des actions de la trace pour mieux focaliser la fouille sur des épisodes pertinents. Dans ce travail nous nous sommes appuyés sur les traces du jeu Tamagocours fournies sous la forme d'un fichier au format `csv`. Il s'agit de rechercher des séquences d'actions signifiantes réalisées par les utilisateurs. Nous présentons dans un premier temps le cadre applicatif des travaux qui serviront à illustrer les concepts présentés par la suite.

2. Data Mining Techniques For Sequence Processing,
`http://liris.cnrs.fr/~crigotti/dmt4sp.html`

| id | date | actionType | group_id | user_id | grpus_id | Codage | message | help | res_id | item_id | resourceType | mode_of_use | resource_title | creationDate | rightsAgreements | res_size | item_size | reason | game_id | level_id | isWon |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | 30/03/2015 11:05:45 | fill Cupboard | 2 | | | | | | | | | | | | | / | | | 2 | 1 | 1 |
| 7 | 30/03/2015 11:06:45 | help Link | 2 | 3 | 2_3 | | | Accès aide en ligne | | | | | | | | / | | | 2 | 1 | 1 |
| 8 | 30/03/2015 11:06:49 | tuto | 2 | 3 | 2_3 | | | | | | | | | | | / | | | 2 | 1 | 1 |
| 12 | 30/03/2015 11:07:42 | showItem CUPBOARD | 2 | 3 | 2_3 | | | | 43 | 339 | book | | Le Grand M | 1913 | Domaine Public | / | 4 | | 2 | 1 | 1 |
| 16 | 30/03/2015 11:07:52 | showItem CUPBOARD | 2 | 3 | 2_3 | | | | 113 | 529 | journal | | Le Figaro | | Domaine public | / | intégrale | | 2 | 1 | 1 |
| 18 | 30/03/2015 11:07:59 | showItem CUPBOARD | 2 | 3 | 2_3 | | | | 43 | 335 | book | | Le Grand M | 1913 | Domaine Public | / | intégrale | | 2 | 1 | 1 |
| 19 | 30/03/2015 11:08:02 | showItem CUPBOARD | 2 | 4 | 2_4 | | | | 108 | 508 | journal | | The Washin | 1983 | Droits d'auteur | / | 5 articles | | 2 | 1 | 1 |
| 20 | 30/03/2015 11:08:19 | addTo Fridge | 2 | 3 | 2_3 | | | | 43 | 335 | book | printedCopies | Le Grand M | 1913 | Domaine Public | / | intégrale | | 2 | 1 | 1 |
| 22 | 30/03/2015 11:08:38 | chat | 2 | 3 | 2_3 | OJ | Quelqu'un sait ce qu'il faut faire?? | | | | | | | | | / | | | 2 | 1 | 1 |
| 23 | 30/03/2015 11:08:39 | showItem FRIDGE | 2 | 4 | 2_4 | | | | 43 | 335 | book | printedCopies | Le Grand M | 1913 | Domaine Public | / | intégrale | | 2 | 1 | 1 |
| 24 | 30/03/2015 11:08:45 | chat | 2 | 6 | 2_6 | OJ | aucune idée! | | | | | | | | | / | | | 2 | 1 | 1 |
| 25 | 30/03/2015 11:08:45 | tuto | 2 | 3 | 2_3 | | | | | | | | | | | / | | | 2 | 1 | 1 |
| 26 | 30/03/2015 11:09:04 | feedTamagoGood | 2 | 3 | 2_3 | | | | 43 | 335 | book | printedCopies | Le Grand M | 1913 | Domaine Public | / | intégrale | | 2 | 1 | 1 |
| 28 | 30/03/2015 11:09:30 | showItem CUPBOARD | 2 | 4 | 2_4 | | | | 108 | 506 | journal | | The Washin | 1983 | Droits d'auteur | / | 1 article | | 2 | 1 | 1 |

TABLE *1 – Un extrait du fichier des traces du jeu Tamagocours. Les attributs sont représentés en colonnes, et la colonne* `actionType` *représente le type d'action réalisé par un utilisateur. Chaque ligne est une action repérée par un identifiant (*`id`*) et par une estampille (*`date`*). Tous les attributs ne sont pas définis pour tous les types d'action (Fuchs, 2017).*

### 3.1 Les traces étudiées

Tamagocours (Sanchez *et al.*, 2015) est un jeu collaboratif destiné à l'apprentissage des règles juridiques auxquelles est soumis l'usage de ressources numériques dans le cadre éducatif. Les utilisateurs sont répartis en groupes de 2 à 4 joueurs et doivent alimenter un «Tamago» avec des ressources pédagogiques. Les utilisateurs peuvent consulter les caractéristiques des ressources et les associer à un mode d'utilisation puis les donner au Tamago pour le «nourrir». Le Tamago est associé à un score qui évolue au fur et à mesure des réussites (ressource autorisée) ou échecs (utilisation d'une ressource hors du cadre légal) des actions des utilisateurs du groupe. Les traces de deux sessions de jeu qui ont eu lieu en 2015 et en 2016 ont été collectées dans deux fichiers au format csv. Elles représentent au total 25 944 lignes pour la session 2015 et 20 752 pour la session 2016, chaque ligne correspondant à une action enregistrée dans le jeu. Les actions sont décrites à l'aide de 24 attributs. Un extrait de ce fichier est montré dans la table 1, et la table 2 contient la liste des types d'actions du jeu.

Dans Tamagocours, une séquence de jeu typique est décrite par la séquence `showItemCUPBOARD,addToFridge,feedTamagoxxxx` qui signifie : l'utilisateur consulte une *ressource* sur l'étagère, range *cette ressource* dans le réfrigérateur puis alimente le tamago avec *cette ressource*. Cette séquence peut être étudiée afin d'observer d'une part son issue qui peut être `xxxx = Good` si l'utilisateur a gagné ou `xxxx = Bad` si l'utilisateur a perdu cette séquence de jeu, et d'autre part les autres actions intercalées dans cette séquence, par exemple utilisation du tutoriel, de l'aide, consultation des autres utilisateurs du groupe par des actions de type «`chat`», etc. Par la suite, nous abrégeons ces deux épisodes respectivement `show,add,Good` et `show,add,Bad`. Il est possible de remarquer que c'est la ressource, associée à un mode d'utilisation, qui relie les trois actions. Néanmoins la fouille avec DMT4SP ne prend pas en compte cette information pour sélectionner les occurrences de motifs, ses choix sont uniquement déterminés par les types d'événement, leur proximité temporelle, et la sémantique d'occurrence minimale que nous décrivons dans la section suivante.

| | |
|---|---|
| addToFridge | ajouter une ressource dans le frigo |
| chat | envoyer un message |
| feedTamagoBad | Nourrir le Tamago avec une bonne |
| feedTamagoGood | ou une mauvaise ressource |
| fillCupboard | Remplissage de l'étagère avec des ressources |
| helpLink | affichage de l'aide |
| removeFromFridge | supprimer une ressource du frigo |
| showItemCUPBOARD | examiner une ressource placée sur l'étagère |
| showItemFRIDGE, | examiner une ressource dans le frigo, |
| showItemLEVEL | examiner une ressource dans le tableau de fin de niveau, |
| showItemTAMAGO | examiner une ressource dans le Tamago. |
| showItemSTOMACH | examiner les ressources dans l'estomac du Tamago. |
| tuto | Consultation du tutoriel |

TABLE *2 – Les différents types d'action utilisateur dans le jeu Tamagocours.*

### 3.2 Extraction d'épisodes séquentiels

DMT4SP est un prototype d'extraction d'épisodes séquentiels et de règles séquentielles à un événement conséquent à partir d'une ou plusieurs séquences d'événements, conformément à une *sémantique d'occurrence minimale* adaptée de (Mannila *et al.*, 1997). DMT4SP prend en entrée une ou plusieurs séquences, un ensemble de paramètres et produit un ensemble d'épisodes séquentiels avec pour chacun d'eux : la fréquence, le nombre de séquences, éventuellement la confiance pour les règles séquentielles, ainsi que, pour chaque occurrence de l'épisode, l'intervalle de temps et le numéro de séquence pour la localiser. Les épisodes séquentiels extraits par DMT4SP sont conformes à la définition suivante :

**Définition 1 (Séquence, événement, type d'événement)**
*Une* ***séquence*** $s = \langle (e_1, t_1), (e_2, t_2), \dots (e_n, t_n) \rangle$ *est une suite ordonnée d'****événements*** *où*
$(e_i, t_i)_{i=1\dots n}$ *est un événement,*
$e_i \in E$ *est un* ***type d'événement****, et* $E$ *est l'ensemble des types d'événements, et*
$t_i \in \mathbb{N}$ *est une estampille associée à* $e_i$ *telle que* $\forall i, t_i \in \mathbb{N}$ *et* $t_i < t_{i+1}$.

**Définition 2 (épisode séquentiel, occurrence minimale, fréquence)**
*Soit* $S = \{s_k\}_k$ *un ensemble de séquences.*
*Un* ***épisode séquentiel*** $p = \langle e_1, e_2, \dots, e_m \rangle, e_i \in E$ *est une séquence de types d'événements de longueur* $m$.
*Une* ***occurrence*** $o$ *de l'épisode séquentiel* $p$ *est une séquence d'estampilles distinctes* $\langle t_1, t_2, \dots t_m \rangle$ *telles que* $\exists k, (e_i, t_i) \in s_k$ *et* $\forall\, i < j \in [1, m], t_i < t_j$.
*Une occurrence* $o$ *de l'épisode* $p$ *est dite* ***minimale*** *si elle n'inclut pas une autre occurrence du même épisode dans* $s_k$*, c'est-à-dire s'il n'existe pas d'occurence* $o' = \langle t'_1, t'_2, \dots, t'_m \rangle$ *telle que* $(t_1 < t'_1$ *et* $t'_m = t_m)$ *ou* $\exists\, i \in [2, m], (t_1 = t'_1$ *et* $t'_i < t_i)$.
*Si* $O = \{o_i\}_i$ *est l'ensemble des occurrences de l'épisode* $p$ *dans* $S$*, la* ***fréquence*** *d'un épisode séquentiel* $p$ *est définie par* $\sigma(p) = |O|$.
*Un épisode séquentiel* $p$ *est* ***fréquent*** *si* $\sigma(p) \geqslant \sigma_{min}$*, où* $\sigma_{min}$ *est le support minimum choisi par l'utilisateur. La fouille retourne un ensemble* $P$ *d'épisodes fréquents tels que* $P = \{p_i\}, \forall i\ \sigma(p_i) \geqslant \sigma_{min}$.

Les définitions ci-dessus appellent plusieurs remarques. Tout d'abord il est possible de remarquer qu'une occurrence d'épisode est définie par des estampilles distinctes, elle ne peut donc pas contenir plusieurs événements distincts ayant la même estampille. Par exemple

dans la séquence suivante :

**Exemple 1 :**

| | | | |
|---|---|---|---|
| | | Good | |
| $S$ | show | add | Good |
| $t_i$ | 1 | 2 | 3 |

DMT4SP ne sélectionne pas l'occurrence $\langle$(show, 1),(add, 2),(Good, 2)$\rangle$ mais sélectionne l'occurrence $\langle$(show, 1),(add, 2),(Good, 3)$\rangle$.

Ensuite, la documentation de DMT4SP, précise que la définition d'une occurrence minimale est adaptée par rapport à la définition de (Mannila *et al.*, 1997) : elle impose que les types d'événements intermédiaires et terminal $e_2$ à $e_m$ doivent se produire «le plus tôt possible». Soient par exemple l'épisode show,add,Good, et trois occurrences de cet épisode resituées ci-dessous dans un extrait de séquence :

**Exemple 2 :**

| | | | | | | |
|---|---|---|---|---|---|---|
| $o_1$ | | show | add | | | Good |
| $o_2$ | show | | add | | | Good |
| $o_3$ | | show | | | add | Good |
| $S$ | show | show | add | remove | add | Good |
| $t_i$ | 1 | 2 | 3 | 4 | 5 | 6 |

Dans cet exemple, la seule occurrence minimale au sens de DMT4SP est $o_1$, qui correspond aux estampilles $\langle$2, 3, 6$\rangle$. DMT4SP fournit donc en sortie l'intervalle [2, 6] pour une seule occurrence. La définition est équivalente à celle de (Mannila *et al.*, 1997) où une occurrence minimale est définie come un intervalle de temps contenant (au moins) une occurrence minimale : définir une occurrence minimale comme un intervalle ne tient pas compte du fait qu'il peut y avoir plusieurs occurrences de l'épisode séquentiel dans cet intervalle. Ceci a une conséquence sur le support (ou fréquence [3]) qui est calculé à partir des intervalles des occurrences minimales et ne reflète pas le nombre réel d'occurrences minimales à l'intérieur de ces intervalles. En réalité, même si la définition de DMT4SP est un peu différente de (Mannila *et al.*, 1997), elle est équivalente, car les occurrences sorties par DMT4SP sont définies par l'intervalle de temps contenant (au moins) une occurrence d'épisode, et la fréquence est définie comme le nombre d'intervalles de temps où l'épisode est trouvé. Les conséquences de ces remarques sur les résultats seront abordées dans la section suivante.

Il existe par ailleurs dans DMT4SP une multitude de contraintes sur les épisodes séquentiels qui s'avèrent utiles afin de limiter les résultats. Dans DMT4SP, deux définitions du support sont prises en compte. La première correspond à celle de la définition 2, c'est-à-dire le nombre d'occurrences minimales trouvées dans toutes les séquences. La deuxième définition est le nombre de séquences qui contiennent au moins une occurrence minimale. Il est possible de spécifier un seuil minimum pour ces deux supports, et seuls les épisodes qui satisfont simultanément ces deux seuils sont sélectionnés. Des contraintes temporelles permettent de limiter l'étalement des épisodes et des événements dans le temps. La fenêtre temporelle est l'intervalle de temps maximal entre le premier et le dernier événement des occurrences d'un épisode. L'intervalle de temps (min/max) entre deux événements consécutifs d'un épisode peut également être précisé. Il est également possible de contraindre la longueur des épisodes (min/max), et leur imposer un préfixe composé d'une séquence d'événements consécutifs par lesquels les épisodes doivent débuter, ou un suffixe constitué d'un seul événement terminal.

3. selon que celui-ci est calculé de façon absolue ou relative

DMT4SP fournit les résultats sous une forme textuelle avec, pour chaque épisode séquentiel satisfaisant les contraintes spécifiées par l'utilisateur : un numéro unique, la liste des types d'événements composant l'épisode, la fréquence (nombre d'occurrences), ainsi que les informations de localisation des occurrences de l'épisode sous la forme d'un intervalle de temps contenant l'occurrence, ainsi que le numéro de séquence.

### 3.3 DISKIT

DISKIT prend en charge les traitements en amont et en aval de la fouille. En pré-traitement, DISKIT construit une ou plusieurs séquences à partir d'une trace en associant chaque type d'action à un type d'événement et inversement au cours du post-traitement. Cette transformation syntaxique est bijective et a pour unique objectif de se conformer au format requis par DMT4SP. Dans la suite, les termes «action» et «événement» seront utilisés comme synonymes de ce fait. Puis DISKIT déclenche la fouille avec les données et paramètres qui lui ont été fournis, récupère les résultats de la fouille et les met en forme afin de les rendre intelligibles. Les résultats sont restitués dans un fichier texte en sortie.

DISKIT effectue par ailleurs d'autres traitements qui ne sont pas pris en charge par DMT4SP. Tout d'abord DISKIT effectue le calcul de la *fermeture* des motifs en post-traitement en s'appuyant sur la fréquence. Deux options permettent, au moment du post-traitement, de sélectionner les épisodes contenant (ou ne contenant pas respectivement) un *pattern* donné sous la forme d'une séquence de types d'événements. Si les types d'événements se retrouvent dans le même ordre sans être nécessairement contigüs dans les épisodes retournés par DMT4SP, ceux ci sont sélectionnés (respectivement éliminés). Dans l'une des expérimentations de la section suivante, nous avons utilisé cette option afin de restreindre l'étude aux épisodes `show,add,Good` et `show,add,Bad` : les trois types d'événements devaient se retrouver dans cet ordre dans les épisodes pour que ces derniers soient sélectionnés. DISKIT permet également de prendre en compte les attributs caractérisant les actions. L'option `split` permet, au moment du pré-traitement, de fractionner une traces en plusieurs séquences en fonction des valeurs des attributs donnés en argument. Ceci permet à DMT4SP de rechercher des épisodes signifiants dans un ensemble de séquences plus petites. Par exemple il est au minimum nécessaire de fractionner la trace en entrée par groupe d'utilisateurs de façon à «isoler» l'analyse des actions des différents groupes, qui travaillent de façon indépendante, dans des séquences séparées. Nous présentons dans la section suivante la mise en œuvre de cette option pour rechercher des occurrences signifiantes d'épisodes séquentiels.

## 4 Prendre en compte le contexte

Une trace est issue de l'observation d'une *activité* et témoigne de l'existence d'actions passées *situées* qui ont été réalisées par des *acteurs* en interaction avec leur *environnement*. Nous nous intéressons ici aux traces laissées par des utilisateurs dans un environnement numérique et issues d'un processus de collecte sous la forme d'éléments *observés* : des actions associées à des éléments de contexte qui apportent des précisions d'ordres différents allant des *objets* manipulés par l'utilisateur par exemple, ou bien d'autres acteurs avec lesquels il interagit, *etc.* Les actions enregistrées dans une trace doivent être associées à un contexte explicite, qui ont été capturées lors du processus de collecte. Néanmoins, il n'est pas possible de «tout» collecter, soit parce que tous les aspects du phénomène étudié ne sont pas observables, soit parce qu'il n'est pas possible de prévoir toutes les utilisations qui pourront être faites des traces en aval. Cependant, dans le cas des traces numériques, le concepteur instrumente l'application

de façon à collecter le plus d'éléments informatifs possibles qui servent de support dans le but d'en faciliter l'interprétation (Champin *et al.*, 2013).

Dans le cadre du jeu Tamagocours, la prise en compte du contexte permet de focaliser l'analyse sur le but poursuivi par les utilisateurs du jeu : le contexte se situe dans le cadre d'équipes constituées de plusieurs utilisateurs (les *acteurs*, qui peuvent être étudiés soit individuellement, soit dans le cadre d'entités «groupe»), dans une séquence de plusieurs jeux (des *mises en situation* de difficultés croissantes) et portant sur des ressources numériques associées à un mode d'utilisation (les *objets*).

Lors de l'exploration des traces du jeu, il s'agit donc de retrouver des épisodes séquentiels qui ont du sens du point de vue du contexte d'apparition des d'actions de la séquence. Concrètement, le contexte est représenté dans les traces par un ensemble d'attributs. Lors de l'exploration des traces par DMT4SP, la proximité temporelle est prépondérante, mais il faut néanmoins tenir compte du fait que les sessions du jeu Tamagocours ont été menées avec plusieurs groupes d'utilisateurs en parallèle, et la trace témoigne de cette organisation : les actions des utilisateurs sont organisées séquentiellement avec le seul critère temporel, et les différentes actions des groupes et des utilisateurs se retrouvent «mélangées» dans la trace. Il est donc obligatoire et indispensable d'organiser les données au moment de la préparation en amont de la fouille pour tenir compte d'une part du désordre dans les données et d'autre part de l'indépendance du travail des différents groupes d'utilisateurs, de sorte que la recherche d'épisodes ait lieu au sein séquences «cohérentes».

Soit par exemple l'extrait suivant de la trace de la session 2015 :

**Exemple 3 :**

| id | 605 | 606 | 608 | 610 | 611 | 612 | 614 | 616 | 618 | 619 |
|---|---|---|---|---|---|---|---|---|---|---|
| action | show | show | show | add | Good | add | Good | add | help | Bad |
| group | 4 | 3 | 3 | 4 | 4 | 3 | 3 | 4 | 4 | 4 |
| user | 13 | 18 | 18 | 12 | 12 | 9 | 9 | 13 | 14 | 13 |
| item | 385 | 363 | 264 | 736 | 736 | 363 | 363 | 385 | | 385 |
| game | 19 | 16 | 16 | 19 | 19 | 16 | 16 | 19 | 19 | 19 |

Si le contexte n'est pas pris en compte, DMT4SP sélectionne les occurrences minimales `(show,608),(add,610),(Good,611)` et `(show,608),(add,610),(Bad,619)`. Les actions `(show,606)` et `(Good,614)` ne peuvent par conséquent pas être sélectionnées dans une occurrence minimale.

Si le numéro de groupe est pris en compte dans le contexte, une occurrence ne peut se situer que dans un même groupe, ce qui revient à analyser les deux séquences suivantes :

**Exemple 4 :**

| id | 606 | 608 | 612 | 614 | 605 | 610 | 611 | 616 | 618 | 619 |
|---|---|---|---|---|---|---|---|---|---|---|
| action | show | show | add | Good | show | add | Good | add | help | Bad |
| group | 3 | 3 | 3 | 3 | 4 | 4 | 4 | 4 | 4 | 4 |
| user | 18 | 18 | 9 | 9 | 13 | 12 | 12 | 13 | 14 | 13 |
| item | 363 | 264 | 363 | 363 | 385 | 736 | 736 | 385 | | 385 |
| game | 16 | 16 | 16 | 16 | 19 | 19 | 19 | 19 | 19 | 19 |

Dans cette situation, l'occurrence `(show,608),(add,612),(Good,614)` peut être sélectionnée pour le groupe 3, et `(show,605),(add,610),(Good,611),(show,605),(add,610),(Bad,619)` pour le groupe 4. Toutefois, aucune de ces occurrences ne porte sur le même numéro d'item (un item est l'association d'une ressource et d'un mode d'utilisation), la seule occurrence qui ait du sens est `(show,605),(add,616), (Bad,619)`, mais celle ci n'est pas considérée comme

minimale par DMT4SP qui ne prend pas en compte le contexte. Il est par conséquent nécessaire de contextualiser davantage les épisodes à l'aide d'autres attributs comme le numéro de jeu car les différents jeux qui se succèdent sont indépendants, mais également sur le numéro d'item. Une meilleure façon d'analyser la trace consisterait donc à rechercher les épisodes dans les séquences suivantes :

**Exemple 5 :**

| | item 385 | | | | item 363 | | | item 264 | item 736 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| id | 605 | 616 | 618 | 619 | 606 | 612 | 614 | 608 | 610 | 611 | 618 |
| action | show | add | help | Bad | show | add | Good | show | add | Good | help |
| group | 4 | 4 | 4 | 4 | 3 | 3 | 3 | 3 | 4 | 4 | 4 |
| user | 13 | 13 | 14 | 13 | 18 | 9 | 9 | 18 | 12 | 12 | 14 |
| item | 385 | 385 | | 385 | 363 | 363 | 363 | 264 | 736 | 736 | |
| game | 19 | 19 | 19 | 19 | 16 | 16 | 16 | 16 | 19 | 19 | 19 |

Dans cet ensemble de séquences, DMT4SP trouverait les deux occurrences minimales : `(show,605),(add,616),(Bad,619)` pour l'item 385 et `(show,606),(add,612),(Bad,614)` pour l'item 363. Il est possible de remarquer dans cette dernière occurrence que des utilisateurs différents ont collaboré pour la réalisation de la séquence de jeu. Il est également possible de remarquer que l'action `help` n'est associée à aucun item, c'est la raison pour laquelle elle apparait dans toutes les séquences du même groupe et du même jeu car il n'est pas possible de *décider* si – et quel – item est concerné par la consultation de l'aide. Ce point sera abordé dans la section 4.1.

### 4.1 Le contexte dans FINEPIO

Nous avons développé un algorithme appelé FINEPIO [4] et qui recherche toutes les occurrences minimales d'un épisode séquentiel en prenant en compte le contexte. FINEPIO prend en entrée une trace, un ou plusieurs épisodes séquentiels dont on souhaite rechercher toutes les occurrences, un contexte sous la forme d'un ensemble d'attributs, et recherche dans la trace toutes les occurrences des épisodes dans le contexte spécifié. On peut remarquer que la fréquence d'un épisode recherché avec FINEPIO peut être supérieure à celle calculée par DMT4SP, car il est possible de trouver plusieurs occurrences d'épisode dans chaque intervalle de temps qui n'est compté qu'une fois par DMT4SP. Les définitions associées à FINEPIO prennent en compte cette particularité dans les définitions suivantes.

**Définition 3 (Séquence, événement)**
*Une* ***séquence*** $s$ *est une suite ordonnée d'événements :*
$s = \langle (e_1, t_1, d_1, c_1), (e_2, t_2, d_2, c_2), \dots (e_n, t_n, d_n, c_n) \rangle$ *où*
*Un* ***événement*** *est un quadruplet* $(e_i, t_i, d_i, c_i)_{i=1\dots n}$ *avec :*
$e_i \in E$, *est le* ***type d'événement****,*
$t_i \in \mathbb{N}$ *est une* ***estampille*** *associée à* $e_i$, $\forall\, i < j \in \mathbb{N}, t_i \leq t_j$,
$d_i \in \mathbb{N}$ ***identifie*** *tout événement de façon unique :* $\forall j \neq i, d_j \neq d_i$,
$C_i = \{(a_i^j, v_i^j)\}$ *est le contexte de l'événement où :*
$A_i = \{a_i^j\}$ *est l'ensemble des attributs associés à l'événement,*
*avec* $A_i \subseteq A$, $A$ *étant l'ensemble de tous les attributs, et*
$v_i^j$, *est la valeur de l'attribut* $a_i^j$ *pour l'occurrence* $i$
*qui prend ses valeurs dans le domaine de valeurs de l'attribut considéré.*

4. FINd EPIsode Occurrences.

**Définition 4 (Contexte, contexte valué)**
*Un contexte $C = \{a^k\}_{k=1...q} \subseteq A$ définit l'ensemble des attributs à prendre en considération dans les événements. C'est un sous-ensemble d'attributs de $A$ dont les valeurs doivent coïncider dans les événements composant les occurrences d'épisodes.*
*Un* ***contexte valué*** *$c_i \subseteq C_i$ est l'ensemble les valeurs prises par les attributs de contexte dans un événement $i$ : c'est un ensemble de couples (attribut, valeur) correspondant à un contexte $C$ pour un événement $i$ :*
$c_i = \{(a_i^j, v_i^j)\}_{i=1...n, j \leq q}$, *tel que* $a_i^j \in C \cap C_i$

D'après cette définition, il est possible, pour un événement $i$, que $C \not\subseteq A_i$, c'est à dire que certains attributs de $C$ ne soient pas définis pour un événement $i$, comme c'était le cas pour l'action `help` dans l'exemple 5.

Dans FINEPIO, la définition d'un épisode séquentiel est identique à celle de DMT4SP, mais la définition des occurrences et des occurrences minimales diffère.

**Définition 5 (épisode séquentiel, occurrence, occurrence minimale)**
*Un* ***épisode séquentiel*** *$p$ est une séquence ordonnée de types d'événements*
$p = \langle e_1, e_2, \ldots, e_m \rangle, e_i \in E$ *de longueur* $m$
*Une* ***occurrence*** $o = \langle (e_1, t_1, d_1, c_1), (e_2, t_2, d_2, c_2), \ldots (e_m, t_m, d_m, c_m) \rangle$ *de l'épisode séquentiel $p$ dans le contexte $C$ est une sous séquence d'une séquence $s$ telle que :*
$(e_i, t_i, d_i, C_i)_{i=1...m} \in s, \forall\, i < j \in [1, m], t_i < t_j$ *et* $\forall\, i \neq j \in [1, m], d_i \neq d_j$ *et*
$\exists (e_k, t_k, d_k, C_k) \in o$ *tel que* $C \subseteq A_k$ *et* $\forall i \neq k, c_i \subseteq c_k$
*Cette dernière contrainte oblige l'existence d'au moins un événement dans l'occurrence qui possède tous les attributs du contexte afin de pouvoir établir l'inclusion.*
*Une occurrence* $o = \langle (e_i, t_i, d_i, c_i) \rangle_{i=1,m}$ *de l'épisode $p$ est* ***minimale***
*si toute occurrence $o'$ de $p$,* $o' = \langle (e'_i, t'_i, d'_i, c'_i) \rangle_{i=1,m}$ *est telle que*
$t_1 < t'_1$ *et* $t_m < t'_m$ *ou* $t_1 > t'_1$ *et* $t_m > t'_m$ *ou* $t_1 = t'_1$ *et* $t_m = t'_m$
*On note $O = \{o_i\}_i$ l'ensemble des occurrences de l'épisode $p$ dans $S$.*
*La* ***fréquence*** *d'un épisode séquentiel $p$ est le nombre d'occurrences minimales de cet épisode dans $S$. Elle est notée $\sigma(p) = |O|$.*

La définition des occurrences d'épisodes impose qu'il existe au moins un événement dans l'occurrence dont le contexte valué contient tous les attributs du contexte, c'est-à-dire un événement $k$ tel que $C \subseteq A_k$. Les contextes valués des autres événements de l'occurrence d'épisode doivent être soit inclus dans ce dernier, soit égal. L'absence de valeur dans un contexte valué ne signifie pas que le contexte est différent, mais qu'il est incomplet, et il est alors *possible* que l'événement correspondant soit pertinent dans un contexte, mais il n'est pas possible de *décider* dans quel contexte précisément. Comme dans l'exemple 5 précédent, une action de type `chat` ou `help` ne comporte pas d'item, mais si ces actions interviennent dans un même groupe et dans un même jeu, elles sont susceptibles de concerner un des items manipulés. La définition des occurrences minimales diffère de celle de (Mannila *et al.*, 1997) où une occurrence minimale est définie comme l'intervalle de temps contenant une occurrence minimale. Dans la définition 5, c'est l'ensemble de toutes les occurrences minimales dans l'intervalle de temps considéré, comme dans l'exemple 2, où les deux occurrences $o_1$ et $o_3$ sont minimales.

## 4.2 Expérimentations

Dans DISKIT, le contexte est pris en compte en répartissant les actions composant la trace dans autant de sous-ensembles qu'il y a de contextes différents. Cela revient à séparer les contextes avant la fouille de façon à assurer que les occurrences d'épisodes sélectionnées par

DMT4SP ont un sens du point de vue de leur contexte. La trace est donc divisée en autant de séquences à explorer qu'il y a de combinaisons de valeurs différentes pour les attributs constituant le contexte. Nous avons expérimenté l'introduction du contexte avec DISKIT afin d'évaluer l'efficacité résultante du point de vue du rappel et de la précision pour des motifs bien identifiés, et leur évolution en fonction de la prise en compte du contexte. Nous avons lancé DISKIT en faisant varier les paramètres de contexte, et nous avons comparé les résultats à ce que nous aurions *idéalement* souhaité trouver dans les résultats et qui a été calculé par FINEPIO. Pour rendre compte de l'efficacité de la stratégie par les indicateurs de rappel et de précision à partir du nombre d'épisodes trouvé par FINEPIO, nous avons focalisés la recherche sur deux épisodes : `show,add,Good` et `show,add,Bad`. Le comptage des occurrences minimales au cours du post-traitement a été réalisé de deux façons différentes par DISKIT. La première utilise la sémantique d'occurrence minimale de DMT4SP et la seconde utilise l'algorithme FINEPIO qui recherche *toutes* les occurrences minimales d'un épisode séquentiel à partir des intervalles fournis par DMT4SP en prenant en compte le contexte. Nous avons exploré les traces de deux sessions de jeu ayant eu lieu en 2015 et en 2016 pour rechercher les deux épisodes `show,add,Good` et `show,add,Bad` et en introduisant progressivement le contexte de ces épisodes : le numéro de groupe, le numéro de jeu, le numéro d'utilisateur et le numéro d'item.

La table 3 contient les principaux paramètres des expériences réalisées. Ils ont été choisis afin de limiter le temps de traitement et le volume de résultats tout en assurant le rappel de toutes les occurrences des épisodes `show,add,Bad` et `show,add,Good`. La longueur minimum et maximum choisie est de 3, la fréquence minimum est de 800, inférieure à celle des deux épisodes `show,add,feed` et rend inutile l'introduction d'une contrainte temporelle pour accélérer le traitement. Le paramètre estampille temporelle permet de calculer les estampilles à l'aide des dates enregistrées dans les actions de la trace (valeur oui) ou à l'aide d'un simple numéro séquentiel (non). Dans ces différentes expériences, nous avons fait varier l'importance de la prise en compte du contexte et noté dans le tableau le nombre de séquences ainsi que le nombre d'événements issus du découpage de la trace. La première expérience a été réalisée sans prise en compte du contexte, la deuxième en introduisant le groupe d'utilisateurs, la troisième expérience en introduisant le groupe et le jeu et la quatrième en introduisant le groupe, le jeu et l'utilisateur. Enfin la cinquième et la sixième ont été réalisées en introduisant le groupe, le jeu et l'item.

| Exp. | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| estampilles temporelles | oui | | | | | non |
| minsup | 800 | | | | | |
| inclusion de patterns | show,add,feedGood show,add,feedBad | | | | | |
| longueur min/max | 3 / 3 | | | | | |
| découpage | | group | group game | user | group game item | |
| nb de séquences | 2 | 152 | 1 177 | 3 359 | 17 557 | 17 557 |
| nb d'événements | 46 696 | 46 696 | 46 696 | 47 909 | 170 884 | 170 884 |

TABLE *3 – Les paramètres utilisés pour l'expérimentation.*

Une caractéristique importante qui rend plus délicate la prise en compte du contexte est que tous les types actions du jeu Tamagocours ne possèdent pas les mêmes attributs. Les actions `chat`, `tuto` et `help` en particulier ne possèdent pas l'attribut `item_id`). Il existe cependant des attributs qui sont partagés par tous les types d'actions (au moins obligatoirement l'identifiant, le type d'action et la date). La définition 5 de FINEPIO prend en compte cette

| | Exp. 1 | Exp. 2 | Exp. 3 | Exp. 4 | Exp. 5 | Exp. 6 |
|---|---|---|---|---|---|---|
| Occurrences d'épisodes retournées par DISKIT | 9276 | 8194 | 7686 | 5381 | 4960 | 4978 |
| retrouvées | 917 | 1821 | 1821 | 4189 | 4958 | 4978 |
| non retrouvées | 4061 | 3157 | 3157 | 789 | 20 | 0 |
| Rappel | 18,42% | 36,58% | 36,58% | 84,15% | 99,60% | 100,00% |
| Précision | 9,89% | 22,22% | 23,69% | 77,85% | 99,96% | 100,00% |

TABLE *4 – Synthèse des résultats obtenus par* DISKIT *pour les occurrences des motifs* `show,add,Good` *et* `show,add,Bad` *en utilisant la sélection des occurrences minimales selon* FINEPIO. *Le nombre d'épisodes trouvés par* FINEPIO *est de 4978.*

| | Exp. 1 | Exp. 2 | Exp. 3 | Exp. 4 | Exp. 5 | Exp. 6 |
|---|---|---|---|---|---|---|
| Occurrences d'épisodes retournées par DMT4SP | 6071 | 5852 | 5696 | 5060 | 4925 | 4942 |
| retrouvées | 650 | 1375 | 1375 | 4171 | 4923 | 4942 |
| non retrouvées | 4328 | 3603 | 3603 | 807 | 55 | 36 |
| Rappel | 13,04% | 27,62% | 27,62% | 83,79% | 98,90% | 99,28% |
| Précision | 10,71% | 23,50% | 24,14% | 82,43% | 99,96% | 100,00% |

TABLE *5 – Synthèse des résultats obtenus par* DISKIT *pour les occurrences des motifs* `show,add,Good` *et* `show,add,Bad`*, lors de l'application de la sémantique d'occurence minimale de* DMT4SP *(le nombre d'épisodes trouvés par* FINEPIO *est de 4978).*

particularité dans la définition d'une occurrence, il ne s'agit pas d'obtenir une identité stricte des contextes dans une occurrence mais une inclusion des contextes, à condition que l'un au moins des événements de l'occurrence possède tous les attributs de contexte. Ceci présente l'avantage l'inclure un contexte de façon plus souple, et d'obtenir des résultats potentiellement plus riches. Mais la contrepartie est que certaines actions doivent être dupliquées dans plusieurs séquences ce qui augmente considérablement le nombre d'événements dans les séquences que DMT4SP doit traiter et a une influence sur le temps de traitement. Dans le cas de Tamagocours, les actions `chat`, `tuto` et `help` possèdent tout de même les deux attributs `group_id` et `game_id`, ce qui limite l'augmentation du nombre d'événements. On peut voir dans la table 3 qu'avec 46 696 actions dans la trace le nombre d'événements générés par le découpage augmente en particulier pour les exp. 5 et 6 pour atteindre 170 884 événements. Dans l'expérimentation réalisée ici, les types d'actions des épisodes `show,add,Good` et `show,add,Bad` possèdent tous les mêmes attributs.

Nous avons expérimenté en post-traitement deux versions de DISKIT conformément aux deux définitions précédentes : la définition 2 de DMT4SP (tableau 5), et la définition 5 de FINEPIO (tableau 4). Chacun d'eux regroupe les résultats obtenus pour les deux épisodes `show,add,Good` et `show,add,Bad` et pour les deux traces. Dans chaque tableau a été reporté le nombre total d'occurrences d'épisodes produits par DISKIT, le nombre d'occurrences d'épisodes pertinents retrouvés parmi ceux qui ont été produits et le nombre d'occurrences qui n'ont pas été retrouvées dans les résultats. Sur ces deux tableaux on peut remarquer que, en l'absence de prise en compte des éléments de contexte, le rappel est très faible et la méthode d'exploration est sans intérêt dans ces conditions car elle s'appuie sur le seul critère de proximité temporelle des événements (Exp. 1) et a très peu de chances de capturer les épisodes signifiants. L'introduction du groupe et du jeu (Exp. 2 et 3) améliore peu le rappel qui reste très faible, du fait des actions mêlées de plusieurs utilisateurs simultanément. Lorsque le numéro d'utilisateur est introduit, le rappel s'améliore de façon significative car le

plus grand nombre d'occurrences d'épisode `show,add,feed` est réalisé par le même utilisateur. Néanmoins, la nature collaborative du jeu fait que parfois plusieurs utilisateurs d'un même groupe contribuent ensemble à la réalisation d'une séquence `show,add,feed` sur un item donné. Par ailleurs un utilisateur peut travailler en même temps sur plusieurs ressources à la fois. Ce critère n'est par conséquent pas fiable. Enfin lorsque un contexte plus précis est introduit avec le numéro de groupe, de jeu et d'item, les séquences sont très contextualisées et le rappel des épisodes devient beaucoup plus important. Le nombre de séquences générées par le découpage avec le contexte devient très important, la taille des séquences devient beaucoup plus faible, ce qui rend le traitement plus rapide et il est alors inutile d'introduire de contraintes temporelles car les intervalles de temps des séquences sont réduits. La légère amélioration observée entre les expériences 5 et 6 vient de la précision des dates. L'unité de temps utilisée dans le jeu est la seconde, et deux actions réalisées de façon très rapide par un utilisateur peuvent par conséquent posséder la même date, même si l'ordre dans lequel elles apparaissent reflète leur instant d'apparition. Il serait nécessaire de re-séquentialiser les actions de la trace pour en tenir compte. La seule différence entre les expériences 5 et 6 est l'utilisation d'un simple numéro séquentiel à la place des dates, et qui reflète l'ordre d'apparition des actions. On obtient une légère amélioration du rappel, et la précision qui devient 100%. Dans le tableau 5, les occurrences d'épisodes non retrouvés dans l'expérience 6 et que l'on retrouve également dans l'expérience 5 sont des occurrences non minimales au sens de DMT4SP, elles sont toutes dues à l'hésitation des utilisateurs qui ont réalisé les séquences `show,add,remove,add,feed`. En effet, dans le jeu Tamagocours, une action `remove` supprime l'effet de l'action `add` qui la précède, et ce n'est que la deuxième action `add` qui rend possible l'action `Good`, comme dans l'exemple 2. Cependant, d'une part cette information n'est pas disponible explicitement et d'autre part elle est spécifique à Tamagocours. Il n'est donc pas possible *a priori* de décider quelle occurrence est la plus pertinente et FINEPIO sélectionne toutes les occurrences. De ce fait, si cette règle était prise en compte pour le comptage des occurrences d'épisodes, la précision de FINEPIO pour l'expérience 6 devrait diminuer légèrement.

### 4.3 Discussion

La préparation des données joue un rôle crucial d'une part pour l'efficacité du traitement et la diminution de la redondance des résultats et d'autre part pour assurer un traitement convenable des données prenant en compte l'organisation des données fournies. La sémantique d'occurrence minimale telle que définie dans DMT4SP est intéressante du point de vue de l'efficacité de la fouille, mais empêche le rappel de toutes les occurrences d'épisode si le contexte n'est pas pris en compte pour préparer convenablement les données. Une conséquence est que l'ensemble des occurrences minimales doivent être recherchées au moment du post-traitement et cette opération augmente légèrement le temps de traitement, mais le fractionnement n'augmente pas significativement le temps de traitement global. Seule la première expérience a un temps de traitement beaucoup plus long car il est plus coûteux de traiter une longue séquence plutôt que plusieurs séquences plus petites.

Dans FINEPIO, deux stratégies ont été mises en oeuvre pour rechercher les occurrences minimales. La première effectue une recherche contextuelle sans découper la trace. Cette méthode doit traiter une très longue séquence et reste assez lente malgré une indexation pour accélèrer le traitement. La deuxième s'appuie sur les attributs présents dans tous les événements (groupe et jeu) pour découper la trace puis effectue une recherche contextuelle dans l'ensemble des traces ainsi découpées. Cette stratégie est beaucoup plus rapide car la recherche est réalisée sur des séquences beaucoup plus petites.

L'intérêt de la contextualisation des épisodes séquentiels est double. Il permet tout d'abord

de fractionner une grande trace en de nombreuses séquences plus petites, rendant inutile l'introduction de contraintes temporelles. L'étape de fouille peut traiter ces séquences de façon efficace et produit moins de résultats redondants. De plus, les épisodes ainsi générés sont cohérents au regard du contexte dans lequel elles portent.

La prise en compte du contexte est réalisée *au sens large*, c'est à dire que si une action ne possède pas de valeur pour l'un au moins des attributs du contexte, alors cette action est rattachée à tous les contextes possédant une valeur pour les attributs considérés. Ce rattachement a pour conséquence une augmentation significative du nombre d'événements que DMT4SP doit traiter. La consultation du *modèle* associé à la trace permet de connaître les attributs associés à chaque type d'action (Fuchs, 2017) et ainsi de vérifier qu'une contrainte de contexte est applicable.

La modélisation proposée sous la forme de contrainte contextuelle est générique et peut s'appliquer à tout type de séquence si les informations présentes dans la trace le permettent. La contextualisation s'avère intéressante pour les traces du jeu Tamagocours, mais il reste encore à vérifier qu'elle l'est également pour d'autres traces et dans d'autres domaines. Elle repose sur l'hypothèse selon laquelle il est possible d'exprimer le contexte sous la forme d'un ensemble d'attributs représentant des notions telles que le sujet de l'action, l'objet de l'action, la phase de l'activité. Il n'est pas évident que toutes les applications et les traces en résultant soient organisés de cette manière, mais on peut raisonnablement penser qu'il existe dans les traces, outre les les types d'actions et leur instant temporel, au moins l'identification des utilisateurs et des objets manipulés dans le jeu. Nous avons observé les traces d'un autre jeu, ClassCraft, dans lequel ces informations sont bien présentes. Mais il reste à bien étudier et comprendre le jeu pour formuler des requêtes et introduire en conséquences des contraintes associées de sorte que l'exploration soit efficace.

Finalement, la façon de prendre en compte le contexte fait que l'on se ramène à une situation proche de la fouille d'une base de séquences, mais l'adéquation de cette méthode pour les traces reste à étudier, du fait de la non prise en compte de la dimension événementielle dans la forme des résultats.

La dimension temporelle est importante essentiellement du point de vue de l'ordre des actions, mais la précision des intervalles de temps entre actions ou entre les première et dernière actions d'un épisode est de moindre importance. Ces expérimentations ont mis en évidence que le contexte est bien plus important que la précision temporelle pour le rappel des occurrences d'épisodes.

## 5 Conclusion

Nous avons proposé une approche pour l'exploration de traces numériques fondée sur les épisodes séquentiels exploitant le contexte des actions pour focaliser l'analyse sur des épisodes signifiants.

La principale perspective à ce travail serait d'intégrer la prise en compte du contexte dans une méthode de fouille de traces à la manière de FINEPIO.

Il reste encore à continuer à développer DISKIT pour être capable de répondre à des requêtes plus complexes que pourraient se poser des utilisateurs d'EIAH. Par exemple, une des options de DISKIT permet notamment de formuler des requêtes sur l'absence de certains types d'événements dans des épisodes, ce qui est utile pour étudier sur l'utilité de certains dispositifs pédagogiques sur les performances des apprenants. D'autres expérimentations restent à mener pour l'éprouver dans d'autres domaines avec des contraintes et des requêtes plus variées. DISKIT ne se limite pas aux traces d'interaction mais à tout type de données symboliques et séquentielles. Nous l'avons utilisé dans des travaux précédents pour des partitions musicales (Fuchs & Cordier, 2016) ayant des caractéristiques différentes des traces.

## Références

AGRAWAL R. & SRIKANT R. (1995). Mining sequential patterns. In *Data Engineering, 1995. Proceedings of the Eleventh International Conference on*, p. 3–14 : IEEE.

BÉCHET N., CELLIER P., CHARNOIS T. & CRÉMILLEUX B. (2014). Fouille de motifs séquentiels pour la découverte de relations entre gènes et maladies rares. *Revue d'Intelligence Artificielle*, **28**(2), 245–270.

CHAMPIN P.-A., MILLE A. & PRIÉ Y. (2013). Vers des traces numériques comme objets informatiques de premier niveau : une approche par les traces modélisées. *Intellectica*, (59), 171–204.

CRAM D. (2010). Découverte interactive et complète de chroniques.

FRAWLEY W. J., PIATETSKY-SHAPIRO G. & MATHEUS C. J. (1992). Knowledge discovery in databases : An overview. *AI Magazine*, **13**(3), 57–70.

FUCHS B. (2017). Assister l'utilisateur à expliciter un modèle de trace avec l'analyse de concepts formels. In C. ROUSSEY, Ed., *Ingénierie des connaissances 2017*, Actes des 28èmes Journées francophones d'Ingénierie des connaissances - IC 2017, p. 151–162.

FUCHS B. & CORDIER A. (2016). Interprétation interactive de connaissances à partir de traces. In N. PERNELLE, Ed., *Ingénierie des connaissances 2016*, Actes des 27e Journées francophones d'Ingénierie des connaissances - IC 2016, p. 167–178.

GENG L. & HAMILTON H. J. (2007). Choosing the right lens : Finding what is interesting in data mining. In *Quality measures in data mining*, p. 3–24. Springer.

MANNILA H., TOIVONEN H. & INKERI VERKAMO A. (1997). Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery*, **1**(3), 259–289.

NANNI M. & RIGOTTI C. (2007). Extracting trees of quantitative serial episodes. In S. DŽEROSKI & J. STRUYF, Eds., *Knowledge Discovery in Inductive Databases : 5th International Workshop, KDID 2006 Berlin, Germany, September 18, 2006 Revised Selected and Invited Papers*, p. 170–188 : Springer Berlin Heidelberg.

PERER A. & WANG F. (2014). Frequence : Interactive mining and visualization of temporal frequent event sequences. In *Proceedings of the 19th International Conference on Intelligent User Interfaces*, IUI '14, p. 153–162 : ACM.

SANCHEZ E., EMIN-MARTINEZ V. & MANDRAN N. (2015). Jeu-game, jeu-play, vers une modélisation du jeu. une étude empirique à partir des traces numériques d'interaction du jeu tamagocours. *Revue STICEF*, **22**.

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

# Modèle de contexte de collaboration : pour qui, pourquoi, comment ?

Siying Li[1], Marie-Hélène Abel[1] et Elsa Negre[2]

[1] Sorbonne universités, Université de technologie de Compiègne, CNRS UMR 7253, HEUDIASYC, 60203 Compiègne cedex, France
{siying.li, marie-helene.abel}@utc.fr
[2] Université Paris-Dauphine, PSL Research University, CNRS UMR 7243, LAMSADE, 75016 Paris, France
elsa.negre@dauphine.fr

**Résumé** : Travailler en collaboration n'est plus une question mais une réalité, la question qui se pose aujourd'hui concerne la mise en œuvre de la collaboration de façon à ce qu'elle soit la plus réussie possible. Il est donc nécessaire de s'interroger sur les critères à prendre en compte pour favoriser son efficacité. Dans le cadre de nos travaux, nous nous intéressons à l'intérêt de prendre en compte le contexte de la collaboration à cette fin. Nous nous sommes interrogées sur sa définition, sa représentation et son exploitation. Cette dernière doit pouvoir se faire à différentes étapes de la collaboration : avant, pendant et après. Dans cet article, nous présentons le modèle du contexte de collaboration que nous avons établi et montrons pourquoi et comment il peut servir à établir une collaboration réussie.

**Mots-clés** : Contexte, Collaboration, Ontologie.

## 1 Introduction

De nos jours, la collaboration entre personnes, organisations et même entre pays favorise le progrès et le développement de la société humaine. Les collaborations réussies permettent de créer de la valeur ajoutée et d'avantager les collaborateurs et le groupe. Cependant, une collaboration réussie n'est pas facile et est conditionnée de différentes manières.

Les collaborateurs doivent faire des ajustements appropriés, face à différentes situations. Et parfois, ils ne savent pas du tout comment améliorer la collaboration. Dans le cadre de notre travail, nous proposons d'utiliser le contexte (informations contextuelles liées à la collaboration) afin d'aider les collaborateurs et d'évaluer leur collaboration.

Malheureusement, le concept de contexte est très dépendant du domaine d'application. Afin de construire un modèle de contexte propre à la collaboration, nous avons analysé les recherches qui ont été faites concernant le contexte.

Ainsi, dans cet article, nous proposons une définition du contexte de collaboration et l'utilisons dans le but de construire un modèle de contexte de collaboration. Cet article est organisé comme suit : les sections 2 et 3 présentent un état de l'art sur les notions de collaboration et contexte. Dans la section 4, nous donnons la définition du contexte de collaboration et sa modélisation, puis nous détaillons un scenario d'utilisation et d'exploitation de ce modèle. Finalement, nous concluons dans la section 5.

## 2 Collaboration

### 2.1 Définition

Le nom *Collaboration* vient du verbe *collaborer*. Il est défini dans le dictionnaire Larousse[1] de la manière suivante : « travailler de concert avec quelqu'un d'autre, l'aider dans ses fonctions; participer avec un ou plusieurs autres à une œuvre commune ». Par ailleurs, Patel et al. (2012) indiquent que la collaboration implique deux ou plusieurs personnes engagées dans une interaction les unes avec les autres, dans un seul épisode ou une série d'épisodes, en travaillant vers des objectifs communs. Selon Suto & Patitad (2015), la collaboration est avant tout un processus de communication mettant l'accent sur les fonctionnalités du transfert des connaissances. Selon Kotlarsky et Oshri (2005), la collaboration est un processus complexe et multidimensionnel.

Notons que les membres d'une collaboration peuvent être des personnes, des groupes de personnes voire même des organisations, par exemple, des collaborations commerciales entre plusieurs entreprises ou institutions. Les membres d'une collaboration se nomment des collaborateurs.

Les actions des collaborateurs pour atteindre l'objectif commun sont des actions collaboratives. Ces dernières sont au final réalisées par des acteurs humains contribuant à la collaboration en tant que membre personne ou bien en tant qu'appartenant au membre groupe ou au membre organisation.

Précisons qu'une action collaborative peut générer un ou plusieurs produits comme résultat et qu'à partir de l'analyse de l'ensemble des actions réalisées, certaines régularités et techniques peuvent être identifiées.

Au final, la définition de collaboration que nous retenons est la suivante :

*Une collaboration est un processus complexe, qui fait intervenir au moins deux collaborateurs et consiste en un ensemble d'actions réalisées par des acteurs humains agissant pour le compte d'un collaborateur afin d'atteindre un objectif commun.*

### 2.2 Discussion

#### 2.2.1 Approches d'évaluation de la collaboration

Différentes approches sont proposées dans la littérature afin d'évaluer une collaboration. Elles peuvent se distinguer selon qu'elles sont quantitatives, qualitatives, multidimensionnelles ou bien mixtes (Baker et al., 2013).

Les approches quantitatives se basent sur la mesure d'indicateurs quantitatifs. Par exemple, afin de mesurer les ressources consacrées aux processus de communication (nommé, l'effort de communication), Hornbæk (2006) utilise des indicateurs tels que le nombre de tours de parole, le nombre de mots prononcés, le nombre d'interruptions, etc.

Bien que très pratique à mettre en œuvre, l'utilisation d'indicateurs pour évaluer une collaboration est difficilement généralisable (Baker et al., 2013). Ainsi l'augmentation ou la diminution de la valeur d'un indicateur peut être liée à différents facteurs et ne pas être univoque (Baker et al., 2013).

Les approches qualitatives se basent sur des indicateurs qualitatifs obtenus au moyen de questionnaires. L'analyse des réponses permet de définir la valeur de l'indicateur à l'aide d'une échelle de cotation à n points (Baker et al., 2013). Par exemple, Savelsbergh et al. (2009) classifient cinq catégories de comportements dans le cadre d'un apprentissage collaboratif en équipe, et demandent aux participants d'évaluer la fréquence de ces comportements au sein de l'équipe, avec une échelle d'évaluation de 5 points allant de (1) jamais, à (5) toujours.

[1] http://www.larousse.fr/dictionnaires/francais/collaborer/17140

Enfin, les approches multidimensionnelles considèrent la collaboration en tenant compte de différentes dimensions (Baker et al., 2013). Ainsi, Schöttle et al. (2014) caractérisent une collaboration par des dimensions telles que la communication, la confiance, l'engagement, le partage des connaissances et l'échange d'informations.

Dans cette mouvance, afin de guider les concepteurs de systèmes de soutien à la collaboration, Briggs et al. (2009) ont établi le Modèle de Collaboration à Sept Couches (en anglais *Seven-Layer Model of Collaboration, SLMC*). Ce modèle définit sept domaines de préoccupation (dimensions) à considérer, un domaine par couche (cf. Tableau 1[2]).

*TABLEAU 1 – Sept domaines de SLMC.*

| **Domaine de préoccupation** | **Description** |
|---|---|
| Objectifs | Un objectif est un état ou un résultat souhaité. |
| Produits | Un produit est un artefact tangible ou intangible ou un résultat produit par le travail du groupe. |
| Activités | Les activités sont des sous-tâches qui, lorsqu'elles sont terminées, donnent les produits qui constituent la réalisation de l'objectif du groupe. |
| Modes de collaboration | Les modes de collaboration sont des régularités observables du comportement et des résultats qui émergent au fil du temps dans le travail d'équipe, par exemple, *générer* (générer plus de concepts dans l'ensemble des idées partagées par le groupe). |
| Techniques | Une technique de collaboration est une procédure réutilisable pour invoquer des interactions utiles entre des personnes travaillant vers un objectif de groupe, par exemple, le *brainstorming*. |
| Outils | Les outils de collaboration sont des artefacts ou des appareils utilisés dans l'exécution d'une opération pour déplacer un groupe vers son objectif. |
| Scripts | Un script est tout ce que les membres de l'équipe se disent et font avec leurs outils pour se rapprocher de l'objectif du groupe. |

Les sept couches sont liées les unes aux autres. Ainsi les collaborateurs utilisent des outils pour mettre en œuvre les techniques leur permettant d'effectuer des activités (ensemble d'actions) pour atteindre un objectif ou un produit. Ce faisant des modes de collaboration sont mis en place et des scripts sont établis. Les modes de collaboration constituent la colonne vertébrale de la collaboration et permettent de la caractériser (Briggs et al., 2009).

Un des avantages présentés par les approches multidimensionnelles concerne l'identification des relations potentielles entre les dimensions distinctives de la collaboration pour l'identification des processus susceptibles d'être améliorés (Baker et al., 2013).

Finalement, les approches mixtes se basent sur les trois types d'approches précédents. Par exemple, Burkhardt et al. (2009) proposent un système de cotation multidimensionnel pour évaluer la qualité de la collaboration dans la conception assistée par des technologies. Ce système comporte sept dimensions et de multiples indicateurs (cf. Tableau 2). Des questionnaires ont été établis et un algorithme permet de calculer la cotation pour chaque dimension en fonction d'un ensemble d'indicateurs (Burkhardt et al., 2009).

[2] Ce tableau est une version simplifiée et récapitulative des sept domaines de préoccupation clés pour les concepteurs de systèmes de soutien à la collaboration de (Briggs et al., 2009), (Locke & Latham., 1990) et (Vreede et al., 2006).

*TABLEAU 2 – Dimensions et indicateurs pour le système de cotation multidimensionnel (Burkhardt et al., 2009).*

| Dimension | Définition | Indicateurs |
|---|---|---|
| Fluidité de la collaboration | Evaluation de la gestion de la communication verbale (tours verbaux), des actions (utilisation d'outils) et de l'orientation de l'attention. | - Fluidité des tours verbaux<br>- Fluidité d'utilisation des outils (stylet, menu)<br>- Cohérence de l'orientation de l'attention |
| Compréhension mutuelle durable | Evaluation des « grounding », processus concernant l'artefact de conception (problème, solutions), les actions des concepteurs et l'état du bureau de réalité augmentée (e.g. fonctions activées). | - Compréhension mutuelle de l'état des problèmes/solutions de conception.<br>- Compréhension mutuelle des actions en cours et des prochaines actions.<br>- Compréhension mutuelle de l'état du système (fonctions actives, documents ouverts) |
| Echange d'informations pour la résolution de problèmes | Evaluation de la mise en commun des idées de conception, le raffinement des idées de conception et la cohérence des idées. | - Génération d'idées de conception (problème, solutions, cas passés, contraintes)<br>- Raffinement des idées de design<br>- Cohérence et suivi des idées |
| Argumentation et atteindre un consensus | Evaluation de l'argumentation et la prise de décision sur un consensus commun ou pas. | - Critiques et argumentation<br>- Vérification de l'adéquation des solutions aux contraintes de conception.<br>- Prise de décision commune |
| Gestion des tâches et du temps | Evaluation de la planification (e.g. répartition des tâches) et de la gestion du temps. | - Planification du travail<br>- Répartition des tâches<br>- Répartition et gestion des tâches interdépendantes<br>- Gestion du temps |
| Orientation coopérative | Evaluation de l'équilibre de la contribution des acteurs dans la conception, la planification et les actions verbales et graphiques. | - Symétrie des contributions verbales<br>- Symétrie d'utilisation des outils graphiques<br>- Symétrie dans la gestion des tâches<br>- Symétrie dans les choix de conception |
| Orientation des tâches individuelles | Evaluation, pour chaque collaborateur, de sa motivation (marques d'intérêt pour la collaboration), son implication (actions) et son attention (orientation de l'attention). | - Acte de motivation et d'encouragement à la motivation des autres<br>- Constance de l'effort mis dans les tâches<br>- Orientation de l'attention en relation avec la tâche de conception |

Les approches mixtes sont multidimensionnelles et utilisent des indicateurs quantitatifs (nombre de réponses) et qualitatifs (réponses positives et négatives). Les approches de ce type sont plus précises, cependant, toute approche pour déterminer la qualité de la collaboration dépend de la façon dont la collaboration est conçue ou définie, et de ce qui est considéré comme étant ses caractéristiques les plus saillantes et les plus pertinentes dans des situations particulières (Baker et al., 2013).

### 2.2.2 Collaboration réussie

Kotlarsky et Oshri (2005) définissent une collaboration réussie comme *le processus par lequel un résultat spécifique, tel qu'un produit ou une performance désirée, est atteint grâce à un effort de groupe*.

Afin de qualifier une collaboration, il est donc nécessaire de considérer des facteurs propres. Kotlarsky et Oshri (2005) utilisent cependant uniquement deux facteurs : le succès du produit et la satisfaction personnelle. De leur côté, Mattessich et Monsey (1992), ont identifié 19 facteurs influençant le succès d'une collaboration. Ils les classent en six groupes : environnement, membres, processus/structure, communication, but et ressources. Quant à Wouters et al. (2017), ils retiennent quatre prérequis à une collaboration réussie :

1. Un objectif partagé entre les parties prenantes impliquées,
2. Une synchronisation des actions,
3. Un échange d'informations, entre les bonnes entités, au bon moment,
4. Une complémentarité entre compétences.

Ainsi selon l'approche d'évaluation de la collaboration retenue, il sera nécessaire de définir les dimensions et les indicateurs adaptés. Les approches multidimensionnelles apparaissent les plus complètes en permettant d'identifier des dimensions critiques de collaboration réussie.

Ainsi le SLMC (Briggs et al., 2009) peut servir de socle à une telle évaluation. Il s'agit cependant de le compléter avec une dimension satisfaction personnelle mise en avant par Kotlarsky et Oshri (2005), une dimension membres, une autre ressource pour tenir compte de l'ensemble des six groupes de (Mattessich & Monsey, 1992). En effet les groupes environnement, processus/structure, communication peuvent être déduits des couches en place. Concernant les prérequis proposés par Wouters et al. (2017), ils sont également bien considérés.

Le premier prérequis intègre les deux premières couches du *SLMC*. L'objectif et le produit du *SLMC* sont implicitement partagés par les acteurs des membres de la collaboration. En comparant les deux premières couches du *SLMC*, on peut obtenir la valeur du facteur succès du produit (Kotlarsky & Oshri, 2005). De plus, l'objectif du *SLMC* est lié au groupe but (Mattessich & Monsey, 1992).

Les deuxième et troisième prérequis sont liés aux couches du *SLMC* concernant les activités (ensemble d'actions), les outils et les techniques, ainsi qu'aux groupes : ressources, processus/structure, environnement, et communication (Mattessich & Monsey, 1992). Ils les précisent cependant. Ainsi les actions doivent être synchrones. Et ils ont besoin d'utiliser diverses ressources (outils, techniques), de suivre certains processus et de satisfaire certaines conditions environnementales. Concernant l'échange d'informations, il doit se faire entre les acteurs pertinents au bon moment par des communications.

Le dernier prérequis n'est pas considéré dans le *SLMC*.

De son côté le *SLMC* tient compte de scripts qui peuvent être considérés comme des traces d'activités. Ces traces sont utiles pour établir des modes de collaboration.

Le transfert de connaissance nécessaire à une collaboration (Murphy et al., 2004) pourra se faire par le biais d'échanges d'informations au sens large. Grâce aux scripts enregistrés, il devient possible d'étudier ces échanges et d'analyser si les actions effectuées traduisent la bonne mise en pratique des connaissances des acteurs. La question qui se pose concerne alors le profil des acteurs de la collaboration : quelles connaissances, compétences possèdent-ils avant la collaboration ? Comment celles-ci évoluent-elles lors de la collaboration ?

Etablir un profil des acteurs de la collaboration est ainsi lié au quatrième prérequis de Wouters et al. (2017) mais également au groupe membres de (Mattessich & Monsey, 1992) et au facteur satisfaction personnelle de (Kotlarsky & Oshri, 2005). Un tel profil permettrait de prendre en considération des caractéristiques complémentaires pouvant intervenir pour favoriser une collaboration réussie. Les échanges d'informations se feront d'autant mieux si les acteurs parlent une même langue, sont familiers avec les mêmes techniques et/ou outils de collaboration, s'apprécient (prise en compte des réseaux).

Pour conclure, qualifier une collaboration de réussie va au-delà de l'atteinte de l'objectif partagé. Différents facteurs doivent être considérés. Ces derniers constituent en quelque sorte le contexte de la collaboration. Ils concernent aussi bien la tâche à réaliser que les moyens au sens large mis en place pour la réaliser : acteurs, outils, techniques, etc.

## 3 Contexte

### 3.1 Définition

Le contexte est une notion complexe (Adomavicius & Jannach, 2014). Selon Kofod-Petersen & Cassens (2006), il est l'élément clé utilisé pour aider des entités intelligentes à comprendre comment les événements dans le monde environnant influencent leur comportement. Il est très mal défini, les définitions de la littérature sont trop dépendantes de leurs propres contextes (Bazire & Brézillon, 2005). Ces dernières dépendent du domaine d'application (Palmisano et al., 2008), citons par exemple :

- Le contexte est l'ensemble des circonstances qui encadrent un événement ou un objet (Bazire & Brézillon, 2005). Cette définition est très utilisée dans le domaine de la psychologie.
- Le contexte représente un ensemble de variables explicites qui modélisent des facteurs contextuels dans le domaine sous-jacent, par exemple, l'heure, le lieu, l'environnement, les périphériques, l'occasion, etc. (Adomavicius & Tuzhilin, 2011). Cette définition vient des systèmes de recommandation.
- Le contexte est l'ensemble des états et paramètres environnementaux qui déterminent le comportement d'une application ou dans lequel un événement se produit et est intéressant pour l'utilisateur (Chen et al., 2000). Cette définition est généralement utilisée dans le domaine de l'informatique contextuelle.

Notons également que dans le domaine de la psychologie, la notion de contexte est souvent utilisée dans le sens suivant : ensemble d'éléments situationnels dans lequel l'objet en cours de traitement est inclus (Bastien, 1998).

Finalement, nous retenons une définition qui semble faire consensus quel que soit le domaine d'application : *le contexte est toute information qui peut être utilisée pour caractériser la situation d'une entité. Une entité est une personne, un lieu ou un objet considéré comme pertinent pour l'interaction entre un utilisateur et une application, y compris l'utilisateur et les applications elles-mêmes* (Dey, 2001).

L'information considérée dans le contexte d'une entité est appelée information contextuelle. Elle permet de représenter les valeurs des facteurs contextuels (Adomavicius & Tuzhilin, 2011). Par exemple, pour le facteur « heure », la valeur : « 17h » est une information contextuelle.

De multiples facteurs de contexte sont mentionnés dans la littérature, tels que le lieu, l'heure, la température ou l'identité de l'utilisateur (Ryan et al., 1999). Brown et al. (1997) ajoutent un autre facteur : les personnes présentes avec les utilisateurs. Citons également l'état des personnes et des groupes (Dey et al., 2001) pour les applications contextuelles, les intentions d'achat (Palmisano et al., 2008) dans le e-commerce, les évènements du moment (Franklin & Flaschbart, 1998) dans les environnements intelligents, l'espace et l'emplacement (Rodden et al., 1998) pour les systèmes mobiles, ou encore l'emplacement de l'utilisateur, les personnes qui l'entourent, les périphériques accessibles ou les modifications apportées au fil du temps (Byun & Cheverst, 2004).

Le regroupement de facteurs permet de définir une dimension mesurable. Par exemple, une dimension temporelle peut regrouper les facteurs : l'heure, les minutes, les secondes, le fuseau horaire, etc. Ainsi le contexte d'une entité devient multidimensionnel.

La définition du contexte d'une entité reste malgré tout dépendante de l'usage que l'on souhaite en faire. Il s'agit donc d'identifier les dimensions adéquates ainsi que les facteurs associés. Ces derniers permettent de définir les caractéristiques d'une entité. De telles caractéristiques peuvent servir non seulement à décrire l'entité à un instant t mais également à inférer des actions possibles ou de nouvelles informations (Kofod-Petersen & Cassens, 2006). Par exemple, dans le cas des prévisions météorologiques, le contexte du jour j est utilisé pour prédire la météo des prochains jours.

En conclusion, après avoir synthétisé ces différents travaux, nous proposons de définir le contexte d'une entité en complétant la définition de Dey (2001) comme suit :

*Le contexte est toute information qui peut être utilisée pour caractériser la situation d'une entité* ***sur une période de temps donnée****. Une entité est une personne, un lieu,* ***un évènement*** *ou un objet considéré comme pertinent pour l'interaction entre un utilisateur et une application, y compris l'utilisateur et les applications elles-mêmes.*

## 3.2 Discussion

Nous nous focalisons sur les informations contextuelles et en particulier sur les dimensions et facteurs associés. Zimmermann et al. (2007) exploitent cinq dimensions fondamentales de contexte d'une entité (présentées dans le Tableau 3) et construisent un modèle de contexte, qui peut être utilisé pour différents types d'entités : entités naturelles, humaines, artificielles et de groupe.

*TABLEAU 3 – Dimensions fondamentales pour le contexte d'une entité (Zimmermann et al., 2007).*

| Dimension | Description |
|---|---|
| Individualité | Contient des propriétés et des attributs décrivant l'entité elle-même |
| Temps | Fournit les coordonnées temporelles de l'entité |
| Localisation | Fournit les coordonnées spatiales de l'entité |
| Activité | Couvre toutes les tâches qui peuvent impliquer cette entité |
| Relation | Représente des informations sur toute relation que l'entité peut établir avec une autre entité |

De leur côté, Sladić & Milosavljević (2017) utilisent les dimensions suivantes : Acteur, Action, Ressource, Moyens, Temps, Lieu et Objectif. Negre (2017), après avoir identifié 10 dimensions de facteurs contextuels : Temps, Individualité, Activité, Relation, Lieu, Objet, Saison, Température, Contexte social et Contexte matériel, n'en retient que 5 dans le cadre des entrepôts de données (Temps, Individualité/Profil Utilisateur, Activité, Relations, Contexte Matériel). De plus, Ferdousi et al. (2017) proposent de décomposer le contexte en 3 familles : le contexte physique (contenant 4 dimensions – temporelle, spatiale, environnementale, équipement), le contexte personnel (et ses 4 dimensions – démographique, sociale, psychophysiologique, cognitive) et le contexte technique (avec 2 dimensions – matérielle et données).

Par comparaison, Zimmermann et al. (2007) ne tiennent pas compte de l'influence des ressources utilisées par une entité au cours d'une activité. Negre (2017) améliore cela en prenant en compte les ressources périphériques via la dimension Contexte Matériel. Sladić & Milosavljević (2017) et Ferdousi et al. (2017), quant à eux, vont plus loin en proposant une dimension Ressource/Données. De notre point de vue, cette dimension du contexte est importante et permettrait d'inférer des informations. En effet, en raison du développement rapide des technologies de l'information et de la communication, les ressources numériques et les métadonnées doivent être prises en compte. Par exemple, une compte-rendu/note de réunion est une ressource au format texte qui peut contenir plusieurs informations contextuelles sur la réunion, telles que l'objectif, l'heure, le lieu, les participants, ...

Une autre approche pour « dimensionner » le contexte est de considérer la difficulté à collecter les informations contextuelles, i.e. les différents niveaux d'abstraction des informations contextuelles de l'entité (Hong et al., 2009). Les facteurs contextuels sont divisés en deux dimensions, de bas niveau et de haut niveau. Les facteurs de bas niveau contiennent les données brutes collectées directement à partir de capteurs physiques (Hong et al., 2009), tandis que les facteurs de haut niveau correspondent aux descriptions agrégées de l'état et de l'environnement de l'entité (Wang & al., 2004). Par exemple, pour un chat entre deux personnes, les facteurs contextuels de bas niveau sont l'heure, l'emplacement, les participants et les enregistrements de discussion. L'objectif de ce chat, quant à lui, est un facteur de haut niveau, qui ne peut pas être obtenu directement/trivialement. L'obtention des facteurs contextuels de

haut niveau est un challenge en soi. Or, ils pourraient apporter une aide précieuse et même avoir une influence déterminante sur l'entité.

Finalement, il apparaît que le dimensionnement du contexte n'est pas figé. La raison principale vient de la complexité et de la dépendance du contexte. Une autre raison est que les informations contextuelles ont une durée de vie et que leur importance évolue au cours du temps. De plus, elles peuvent être dynamiques ou statiques, ce qui dépend du moment. Par exemple, pendant un même mois, l'âge d'une personne est relativement statique alors que si cette personne déménage durant ce mois, son adresse est dynamique. Par conséquent, l'adresse est plus « importante » d'un point de vue contextuel que l'âge pour ce mois donné.

## 4 Contexte de collaboration

Lorsqu'on parle de contexte, il est nécessairement associé à une entité : le contexte de quoi? Comme nous l'avons vu en section 3, le contexte précise les dimensions contextuelles de l'entité qu'il décrit. Dans ce qui suit nous étudions les dimensions contextuelles à prendre en considération lorsque l'entité à décrire est une collaboration. Nous parlerons de contexte de collaboration.

Cette section traitera du contexte de collaboration à partir de trois aspects, définition, modélisation et exploitation. La section 4.1 discutera de sa définition concernant les concepts apparentés présentés dans les sections 2 et 3. Ensuite, la section 4.2 montrera la modélisation correspondante du contexte de collaboration. Enfin la section 4.3 utilisera et exploitera le modèle à différentes étapes de la collaboration : avant, pendant et après.

### 4.1 Définition

Si on parle de contexte de collaboration, littéralement, l'entité de la définition retenue dans la section 3.1 est la collaboration. Son contexte contient des informations caractéristiques qui sont fortement dynamiques et qui évoluent au cours du temps.

Précisément la définition devient :

*Le contexte **de collaboration** est toute information qui peut être utilisée pour caractériser la situation **de collaboration** sur une période de temps donnée. **Ici, la collaboration est un évènement** considéré comme pertinent pour l'interaction entre un utilisateur et une application, y compris l'utilisateur et les applications elles-mêmes.*

Précisons que les utilisateurs sont des acteurs humains membres de la collaboration. Quant aux applications, elles peuvent être de tout type.

En s'appuyant sur la définition de la collaboration donnée et discutée en section 2, les informations contextuelles de collaboration incluent les membres de la collaboration, l'objectif commun, des activités (ensemble des actions collaboratives) faites, la période du temps de collaboration, des outils, techniques, scripts et modes de collaboration utilisés et des produits. Toutes ces informations sont regroupées en facteurs ou dimensions afin de construire un modèle de contexte de collaboration.

### 4.2 Modélisation

Nous nous appuyons sur la définition du contexte de collaboration ci-dessus pour établir son modèle et retenons principalement trois concepts liés les uns aux autres : information, facteur et dimension contextuelle.

Comme nous l'avons discuté dans la section 3.1, l'information contextuelle est la valeur d'un facteur contextuel. Les facteurs peuvent être regroupés en dimensions contextuelles pour décrire la situation d'une entité. Ces trois concepts contextuels peuvent constituer une architecture générale du modèle de contexte (cf. Figure 1). Les dimensions fondamentales

définies dans (Zimmermann et al., 2007) respectent cette architecture et peuvent être reprises pour différents types d'entités.

La construction du modèle de contexte de collaboration peut donc se baser sur cette dernière. Il s'agit alors de définir les dimensions, leurs facteurs et leurs domaines de valeurs en fonction du pourquoi nous voulons l'utiliser : pouvoir établir, analyser, et mesurer la réussite d'une collaboration.

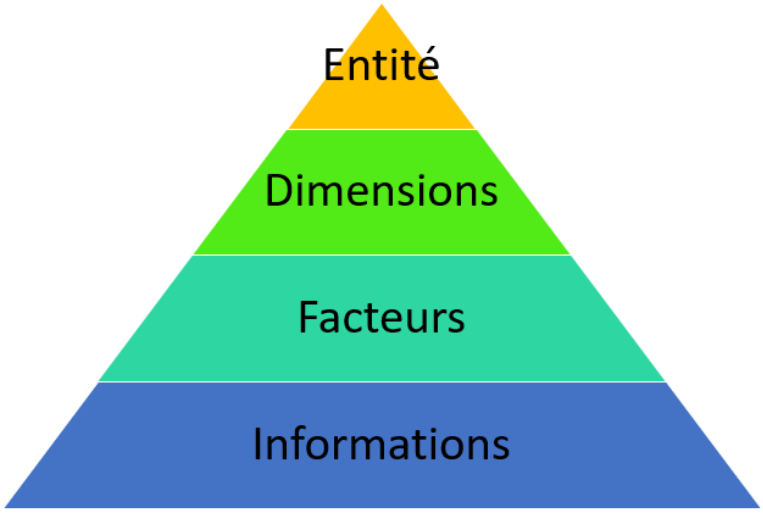



FIGURE 1 – L'architecture du modèle de contexte.

Côté collaboration, en reprenant la définition et la discussion de la section 2, les informations caractérisées pour refléter la situation d'une collaboration sont les suivantes : collaborateur, activité (ensemble d'actions), acteurs humains, objectif, satisfaction personnelle, modes de collaboration, techniques, outils, scripts et produits.

Les collaborateurs, en tant que membres d'une collaboration, peuvent se distinguer selon qu'ils sont individuels (personne) ou collectif (groupe ou organisation). Les collaborateurs partagent le même objectif. Les acteurs humains, effectuent des activités (ensemble d'actions) et utilisent des outils, des techniques pour le compte du collaborateur, les scripts et les modes de collaboration peuvent alors être observés. Les activités réalisées produisent des résultats (produits) pour atteindre le résultat final : l'objectif partagé. Un degré de satisfaction personnelle est émis par chaque acteur humain membre de la collaboration.

Côté contexte, en reprenant la définition et la discussion de la section 3, huit dimensions sont retenues : Temps, Localisation, Activité, Relation, Collaborateur, Ressource, Satisfaction personnelle et Objectif. Les quatre premières sont directement inspirées des travaux Zimmerman et al (2007). Les quatre dernières permettent d'identifier les membres de la collaboration ainsi que de préciser les ressources qu'ils vont exploiter dans le cadre de la collaboration, l'objectif commun visé et la satisfaction personnelle ressentie dans le cadre de la collaboration (cf. Figure 2).

Dans ce modèle nous avons clairement défini le concept de collaboration qui reprend en quelque sorte la dimension individualité de Zimmerman. L'objectif est défini en tant que dimension afin de permettre de le lier aux produits qui seront créés ou utilisés par les activités effectuées par les collaborateurs. Ces derniers sont donc modélisés en tant que tels et une dimension leur est donc dédiée. Des ressources sont utilisées pour effectuer ces activités d'où

la création de cette dimension. Enfin, le fait d'avoir créé le concept *collaboration*, nous permet de prendre en considération la dimension de relation sous la forme d'une relation.

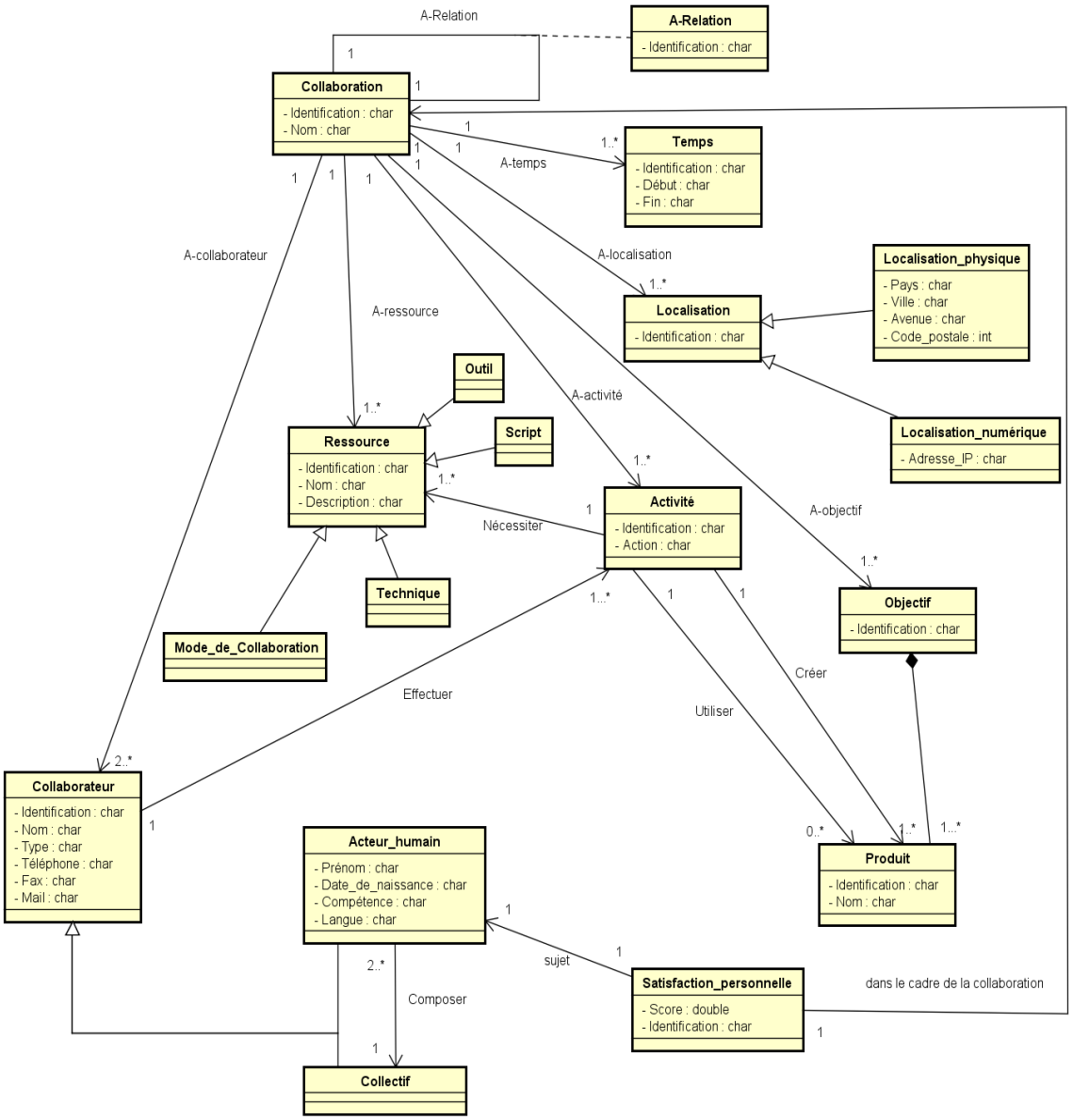



FIGURE 2 – Modèle de contexte de collaboration.

Le tableau 4 présente une synthèse des dimensions retenues avec leurs facteurs.

*TABLEAU 4 – Résumé des dimensions et facteurs du modèle de contexte de collaboration.*

| Dimension | Description | Facteurs |
|---|---|---|
| Temps | Cette dimension fournit des facteurs temporels d'une collaboration. | Identification, Début et Fin. |
| Localisation | Cette dimension contient des facteurs spatiaux d'une collaboration. | - Localisation physique : Identification, Pays, Ville, Avenue/rue et Code_postal.<br>- Localisation numérique : Identification et Adresse_IP. |
| Activité | Cette dimension couvre toutes les actions réalisées par des acteurs humains dans le cadre d'une collaboration. | Identification, Action, au moins un produit(Créer), produit(Utiliser) et au moins une ressource(Nécessiter). |
| Relation | Cette dimension représente des facteurs sur toutes les relations que la collaboration peut établir avec une autre collaboration. | Identification et deux collaborations. |
| Collaborateur | Cette dimension contient tous les facteurs sur les membres de la collaboration. Un collaborateur soit individuel (un acteur humain) ou collectif. | - Acteur humain : Identification, Nom, Type, Téléphone, Fax, Mail, Prénom, Date_de_naissance, Compétence, Langue et au moins une Activité.<br>- Collectif : Identification, Nom, Type, Téléphone, Fax, Mail, au moins deux Acteur humains et au moins une Activité. |
| Ressource | Cette dimension contient des facteurs concernant les ressources qui peuvent être utilisées afin de réaliser des actions collaboratives. Les ressources peuvent être une technique, un outil, un mode de collaboration ou un script. | - Technique : Identification, Nom et Description.<br>- Outil : Identification, Nom et Description.<br>- Mode de collaboration : Identification, Nom et Description.<br>- Script : Identification, Nom et Description. |
| Satisfaction personnelle | Cette dimension précise un degré de satisfaction d'un membre d'une collaboration pour cette dernière. | Identification et Score. |
| Objectif | Cette dimension décrit l'objectif partagé d'une collaboration et il est lié à des produits créés ou utilisés par les activités. | Identification et au moins un produit. |

Notons que les facteurs au sein des dimensions peuvent se traduire sous la forme de simples attributs/valeurs ou bien sous la forme de propriétés. Le facteur *Mode de collaboration* consiste en l'enregistrement des traces d'interaction entre collaborateurs. Le fait d'avoir modélisé le concept de collaborateur permet également de prendre en considération le profil de ce dernier qu'il se décline sous la forme d'un acteur humain ou bien d'un collectif (groupes d'acteurs humains).

### 4.3 Scenario d'utilisation et d'exploitation

Dans cette section nous discutons de l'intérêt du modèle du contexte de collaboration établi dans la perspective d'une collaboration réussie. Ce modèle peut ainsi servir en prélude à une collaboration, à guider une collaboration ou encore à faire le bilan d'une collaboration. Nous illustrons notre propos à partir d'un scénario.

### 4.3.3 Prélude d'une collaboration

Afin d'établir une collaboration, il est nécessaire d'identifier ou de définir les caractéristiques de cette dernière. Le modèle du contexte de collaboration établi, par l'intermédiaire du renseignement des dimensions et des facteurs retenus, peut servir de support pour cette identification. Une collaboration se prépare : Quel est l'objectif de la collaboration ? Quelle est la durée prévue pour cette dernière ? Qui seront les collaborateurs ? Quelles seront les tâches/activités qui leur seront assignées ? Une collaboration réussie dépend de ces identifications. Dans la mesure où l'objectif est bien défini, il devient plus facile d'identifier les compétences nécessaires à sa réalisation et de fait les collaborateurs à recruter présentant de telles compétences. Ces derniers identifiés, il s'agit de sélectionner ceux qui semblent le plus à même de bien travailler ensemble : ceux qui utilisent une même langue de travail, les mêmes outils et techniques de travail, qui ont déjà collaborés, etc.

### 4.3.4 Guidage de la collaboration

Au gré de la collaboration, certains indicateurs peuvent servir à identifier des points de blocage. Les activités prévues sont-elles réalisées dans les temps ? Les interactions entre collaborateurs se font-elles régulièrement ? Ces indicateurs peuvent être obtenus au moyen de l'observation/étude des traces d'interaction enregistrées à partir du modèle. A partir de ces indicateurs, il devient possible de mettre en place une stratégie pour remédier aux problèmes qui causent la mauvaise valeur de ces derniers.

Dans notre scénario, le fait d'observer qu'il n'y a pas eu d'interaction entre les collaborateurs affectés à la création d'un produit depuis 7 jours et que la livraison est prévue dans 3 jours alors qu'aucun élément ne permet de voir si le produit est finalisé, peut permettre au leader de la collaboration de contacter les collaborateurs à des fins d'investigation.

### 4.3.5 Bilan de la collaboration

Lorsque la collaboration est terminée, il est toujours intéressant de tirer des enseignements de son déroulement et des résultats obtenus. Il est assez aisé de mesurer si l'objectif est atteint mais mesurer la réussite de la collaboration est plus délicat. Effectuer cette mesure peut se faire en étudiant le bilan de la collaboration, bilan qui reprend l'ensemble des activités menées. Ces activités permettent de lier ressources, produits et collaborateurs dans le cadre d'une tâche de collaboration. Il est possible de mesurer la quantité, la fréquence des actions effectuées entre quels collaborateurs et avec quels outils. A partir de ces mesures et en les croisant avec le profil des collaborateurs, il devient possible d'avancer des éléments d'analyse et d'en tirer des conclusions afin de prévoir de nouvelles collaborations.

Dans notre scénario, les interactions régulières entre les collaborateurs au moyen des ressources établies et l'atteinte de l'objectif dans les temps sont des indicateurs d'une collaboration réussie. A contrario, bien que l'objectif ait été atteint, l'observation d'un nombre très faible d'interactions peut interpeller : la complémentarité des compétences des collaborateurs était-elle adaptée ? Y a-t'il eu un problème avec les ressources choisies (outils techniques) qui ne seraient pas appropriées ?

Au final, l'utilisation de notre modèle du contexte de collaboration permet, à partir d'éléments tangibles, d'évaluer une collaboration selon plusieurs dimensions, telles que les ressources, les collaborateurs et les activités. Cette évaluation vise à caractériser une collaboration avec un degré de réussite.

## 5 Conclusion

Dans cet article, nous nous intéressons à la notion de contexte de collaboration qui permet de qualifier la collaboration. A partir des travaux de la littérature sur la collaboration et le contexte, nous avons établi une définition du contexte de collaboration. Nous en avons alors

construit un modèle à partir d'une architecture de modèle de contexte d'une entité, présenté sous forme de pyramide. Ce modèle décrit la collaboration selon une approche multidimensionnelle. Nous avons finalement expliqué pourquoi et comment notre modèle pouvait servir la mise en place d'une collaboration (son prélude), sa mise en œuvre (soutien au déroulement, guidage) et l'analyse de son bilan.

Nos perspectives de recherche incluent le développement de ce modèle sous la forme d'une ontologie et son exploitation au sein d'un environnement numérique, ainsi que celui d'un système de recommandation. Grâce aux traces et données enregistrées au sein de l'environnement numérique, le système visé aura la capacité de quantifier et de qualifier la collaboration. Des recommandations pourraient alors être émises de manière statique lors de la phase de prélude de la collaboration et du bilan. Elles pourraient se faire également de manière dynamique lors du déroulement de la collaboration. Un travail d'approfondissement mériterait finalement d'être réalisé sur les relations qu'entretient un tel modèle avec celui du contexte de l'utilisateur (ou contexte d'une personne/acteur).

## Références

ADOMAVICIUS G. & JANNACH D. (2014). Preface to the special issue on context-aware recommender systems. *User Modeling and User-Adapted Interaction*, volume 24, p. 1-5: Springer.

ADOMAVICIUS G. & TUZHILIN A. (2011). Context-Aware Recommender Systems. *Recommender Systems Handbook*, p. 217.

BAKER M., DÉTIENNE F. & BURKHARDT J.-M. (2013). Quality of collaboration in design: articulating multiple dimensions and viewpoints. In *1st Interdisciplinary Innovation Conference*, Telecom ParisTech.

BASTIEN C. (1998). Contexte et situation. In Houdé O., Kayser D., Koenig O., Proust J. & Rastier F., *Dictionnaire des Sciences Cognitives*. Paris: PUF.

BAZIRE M. & BRÉZILLON P. (2005). Understanding context before using it. In *International and Interdisciplinary Conference on Modeling and Using Context*, p. 29-40: Springer, Berlin, Heidelberg.

BRIGGS R., KOLFSCHOTEN G., VREEDE G., ALBRECHT C., DEAN D. & LUKOSCH S. (2009). A seven-layer model of collaboration: Separation of concerns for designers of collaboration systems. *ICIS 2009 Proceedings*, p.26.

BROWN P., BOVEY J. & CHEN X. (1997). Context-aware applications: from the laboratory to the marketplace. *IEEE personal communications*, volume 4, p. 58–64.

BURKHARDT J.-M., DÉTIENNE F., HÉBERT A.-M. & PERRON L. (2009). Assessing the "quality of collaboration" in technology-mediated design situations with several dimensions. In *IFIP Conference on Human-Computer Interaction*, p. 157-160: Springer, Berlin, Heidelberg.

BYUN H. & CHEVERST K. (2004). Utilizing context history to provide dynamic adaptations. *Applied Artificial Intelligence*, volume 18, p. 533-548: Taylor & Francis.

CHEN G. & KOTZ D. (2000). *A survey of context-aware mobile computing research*. Technical Report. Dartmouth College, Hanover, NH, USA.

DEY A. (2001). Understanding and using context. *Personal and ubiquitous computing*, volume 5, number 1, p. 4–7: Springer-Verlag.

DEY A., ABOWD G. & SALBER D. (2001). A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. *Human-computer interaction*, volume 16, number 2, p. 97–166: L. Erlbaum Associates Inc.

FERDOUSI Z., NEGRE E. & COLAZZO D. (2017). Context Factors in Context-Aware Recommender System. *AISR 2017 Atelier interdisciplinaire sur les systèmes de recommandation*. Paris, France.

FRANKLIN D. & FLASCHBART J. (1998). All gadget and no representation makes jack a dull environment. In *Proceedings of the AAAI 1998 Spring Symposium on Intelligent Environments*, p.155–160.

HORNBÆK K. (2006). Current practice in measuring usability: Challenges to usability studies and research. *International journal of human-computer studies*, volume 64(2), p. 79-102.

KOFOD-PETERSEN A. & CASSENS J. (2006). Using activity theory to model context awareness. In *Modeling and Retrieval of Context*, p. 1–17: Springer, Berlin, Heidelberg.

KOTLARSKY J. & OSHRI I. (2005). Social ties, knowledge sharing and successful collaboration in globally distributed system development projects. *European Journal of Information Systems*, volume 14(1), p. 37-48.

LOCKE E.-A. & LATHAM G.-P. (1990). *A theory of goal setting & task performance*. Englewood Cliffs, NJ, US: Prentice-Hall, Inc.

MATTESSICH P.-W. & MONSEY B.-R. (1992). Collaboration: what makes it work. Amherst H. Wilder Foundation, 919 Lafond, St. Paul, MN 55104.

MURPHY F., STAPLETON L. & SMITH D. (2004). Tacit Knowledge and human centred systems: The key to managing the social impact of technology. International Multitrack Conference of Advances in Control systems, University of Vienna (TUWien), Austria.

NEGRE E. (2017). Prise en compte du contexte dans les systèmes de recommandations de requêtes OLAP. *EDA 2017*, p. 1-10.

PALMISANO C., TUZHILIN A. & GORGOGLIONE M. (2008). Using context to improve predictive modeling of customers in personalization applications. *IEEE transactions on knowledge and data engineering*, volume 20, number 11, p. 1535–1549.

PATEL H., PETTITT M. & WILSON J.-R. (2012). Factors of collaborative working: A framework for a collaboration model. *Applied ergonomics*, volume 43, number 1, p. 1–26.

RODDEN T., CHEVERST K., DAVIES K. & DIX A. (1998). Exploiting context in hci design for mobile systems. In *Workshop on human computer interaction with mobile devices*, p. 21–22: Glasgow.

RYAN N., PASCOE J. & MORSE D. (1999). Enhanced reality fieldwork: the context aware archaeological assistant. *Bar International Series*, volume 750, p. 269–274.

SAVELSBERGH C.-M., VAN DER HEIJDEN B.-I. & POELL R.-F. (2009). The development and empirical validation of a multidimensional measurement instrument for team learning behaviors. *Small Group Research*, volume 40(5), p. 578-607.

SCHÖTTLE A., HAGHSHENO S. & GEHBAUER F. (2014). Defining cooperation and collaboration in the context of lean construction. In *Proc. 22nd Ann. Conf. of the Int'l Group for Lean Construction*, p. 1269-1280.

SLADIĆ G. & MILOSAVLJEVIĆ B. (2017). Context-Aware Access Control for IoT Driven Processes. In *the 8th PSU-UNS International Conference on Engineering and Technology (ICET-2017),* Akdeniz University, Antalya, Turkey.

SUTO H. & PATITAD P. (2015). A representation model of collaboration in design process. In *Control Conference (ASCC), 2015 10th Asian*, p. 1–5: IEEE.

DE VREEDE G.-J. D., KOLFSCHOTEN G.-L. & BRIGGS R.-O. (2006). Thinklets: a collaboration engineering pattern language. *International Journal of Computer Applications in Technology*, volume 25, p. 140–154: Inderscience Publishers.

WANG X., DONG J.-S., CHIN C.-Y., HETTIARACHCHI S. & ZHANG D. (2004). Semantic space: An infrastructure for smart spaces. *IEEE Pervasive computing*, volume 3, p. 32–39.

WOUTERS L., CREFF S., BELLA E.-E. & KOUDRI A. (2017). Collaborative systems engineering: Issues & challenges. In *Computer Supported Cooperative Work in Design (CSCWD), 2017 IEEE 21st International Conference on*, p. 486–491: IEEE.

HONG J.-Y., SUH E.-H. & KIM S.-J. (2009). Context-aware systems: A literature review and classification. *Expert Systems with applications*, volume 36, p. 8509–8522: Elsevier.

ZIMMERMANN A., LORENZ A. & OPPERMANN R. (2007). An operational definition of context. In *International and Interdisciplinary Conference on Modeling and Using Context*, p. 558–571: Springer, Berlin, Heidelberg.

# Démonstrations et posters

# Projet VisaTM : l'interconnexion OpenMinTeD – AgroPortal – ISTEX, un exemple de service de *Text et Data Mining* pour les scientifiques français

Fabienne Kettani,[1] Stéphane Schneider,[1] Sophie Aubin,[4] Robert Bossy,[3] Claire François,[1] Clément Jonquet,[2] Andon Tchechmedjiev,[2] Anne Toulet,[2] et Claire Nédellec[3]
[1]CNRS-INIST (Institut de l'Information scientifique et Technique), Vandœuvre-lès-Nancy, France
fabienne.kettani@inist.fr
[2]Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, CNRS & Univ. de Montpellier, France
[3]Unité MaIAGE (Mathématiques et Informatique appliquées du Génome à l'environnement), INRA, Jouy-en-Josas, France,
claire.nedellec@inra.fr
[3]Unité DIST (Direction Information Scientifique et Technique), INRA, Versailles, France,

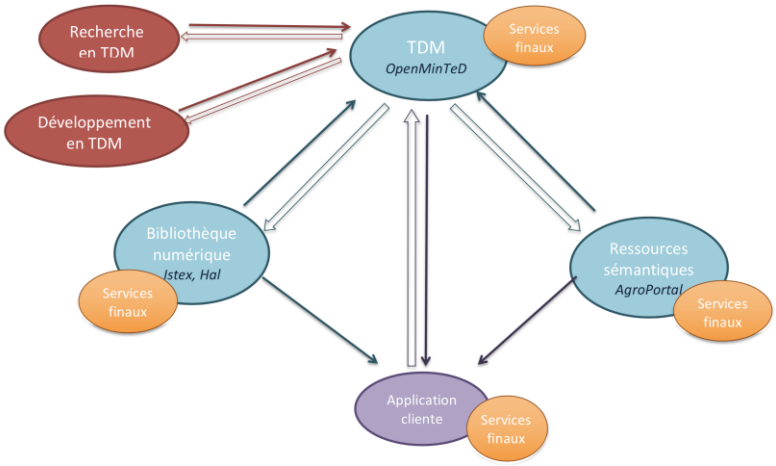
**Mots-clés** : Fouille de texte et de données, ontologies et ressources sémantiques, corpus de données scientifiques, analyse sémantique, OpenMinTeD, ISTEX, AgroPortal.

## 1 Présentation du projet VisaTM

La création d'une offre de service en fouille de texte et de données – TDM (*Text and Data Mining*) – à destination des scientifiques se pose dans un contexte évolutif sur le plan légal,[1] organisationnel et scientifique. Les progrès récents des méthodes d'analyse textuelle ouvrent la voie à l'intégration d'informations extraites des corpus de textes avec des connaissances et données externes souvent publiées sur le Web (e.g, *Linked Open Data*), grâce à l'analyse sémantique basée sur des terminologies et des ontologies dans des domaines spécialisés.

Le projet VisaTM, de la BSN (Bibliothèque Scientifique Numérique),[2] rassemble dans un partenariat trois institutions mettant en synergie leurs complémentarités pour définir les contours d'une infrastructure de services avancés pour la fouille de texte en France : l'INRA, partenaire du projet H2020 d'infrastructure de fouille de texte OpenMinTeD (http://openminted.eu), le CNRS-INIST porteur de l'infrastructure ISTEX (www.istex.fr) et le LIRMM porteur du projet AgroPortal (http://agroportal.lirmm.fr).

L'objectif du projet VisaTM est d'étudier les conditions de production de services de TDM à haute valeur ajoutée basée sur l'analyse sémantique de contenu pour les scientifiques en France. Il doit permettre aussi d'imaginer et de décrire un dispositif technique et humain d'interconnexion d'une instance d'OpenMinTeD avec des bibliothèques numériques et des ressources sémantiques et mettre ainsi en évidence l'opportunité de l'implantation d'une telle infrastructure en France (Figure 1).




Le projet se décline en trois volets distincts : (i) une *étude* d'opportunité

[1] Par exemple, les articles 30 et 38 de la Loi République Numérique sur la libération des articles scientifiques et le TDM.

[2] La BSN (www.bibliothequescientifiquenumerique.fr) est une initiative du MESRI, dans le cadre de la stratégie « Open science », pour l'accès et la diffusion de l'information scientifique, devenue en 2018 le Comité pour la Science Ouverte (CoSO). Ce travail est financé par le programme BSN-10 en 2017-2018.

visant à situer le contexte actuel et les besoins des différents acteurs en TDM afin de proposer une infrastructure adaptée et optimale ; (ii) l'élaboration de trois *applications pilotes* destinées à donner des exemples concrets d'utilisation des interconnexions de la plateforme OpenMinTeD avec des bibliothèques numériques et portails de ressources sémantiques ; (iii) un volet *conception* ciblant les verrous techniques de ces interconnexions. La bibliothèque numérique ISTEX et le portail de ressources sémantiques AgroPortal sont utilisés ici comme preuve de concept pour une utilisation intégrée de la plateforme OpenMinTeD.

## 2 Description des 3 composants : la plateforme OpenMinTeD, ISTEX et AgroPortal

**OpenMinTeD :** L'offre de services TDM s'est développée de longue date sur des plateformes basées sur l'assemblage de composants réutilisables, combinables et adaptables à différentes tâches. Le partage et l'interopérabilité des composants issus de ces plateformes sont au cœur de la création de l'infrastructure de TDM européenne OpenMinTeD. Elle met en effet à disposition un environnement complet en accès ouvert incluant non seulement une bibliothèque complète de composants TDM, mais aussi la possibilité de composition de *workflows*[3] et leur exécution sur un *cloud.* Elle est à destination de diffèrent types de public – spécialistes, développeurs, utilisateurs – dans plusieurs domaines pilotes dont les sciences humaines et sociales, l'agriculture et les sciences de la vie. L'apprentissage automatique et l'utilisation de corpus variés et de ressources sémantiques spécialisées sont des éléments clefs de l'adaptation des applications aux besoins. Ainsi, via ses APIs, OpenMinTeD est connectable de façon standardisée à d'autres infrastructures dont, d'ores et déjà, des infrastructures européennes de contenus, articles et ressources *Open Access*, telles que Open AIRE, ou CORE. Dans le projet VisaTM, un de nos objectifs est d'établir des interconnexions supplémentaires avec les plateformes ISTEX et AgroPortal.

**ISTEX :** les composants TDM analysent des contenus textuels sous forme de corpus adaptés aux thématiques des chercheurs. ISTEX est une bibliothèque numérique de grande envergure (environ 21 millions d'objets à ce jour), regroupant les archives scientifiques acquises sous licence nationale.[4] Le projet comporte deux dimensions : l'acquisition de collections numériques rétrospectives et la mise au point d'une plateforme les rendant facilement accessibles, exploitables et interrogeables automatiquement. Dans le projet VisaTM nous travaillons à l'intégration d'ISTEX comme source de contenu dans la plateforme OpenMinTeD. Un utilisateur aura ainsi la capacité de constituer un corpus sur ISTEX par le biais d'une recherche, de l'enregistrer et de renseigner les métadonnées décrivant le corpus sur la plateforme OpenMinTeD pour l'utiliser. L'implémentation de l'API OpenMinTeD permet des opérations telles que la recherche par mots clefs, le téléchargement des métadonnées et le téléchargement du contenu. Les métadonnées des corpus déclarés sur OpenMinTeD doivent être converties au format OMTD-SHARE de métadonnées pivot supporté par la plateforme vers lequel les descriptions des ressources ISTEX sont alignées. En outre, l'accès aux ressources ISTEX étant limité à l'enseignement supérieur français, l'interconnexion doit gérer les aspects de restrictions d'accès.

**AgroPortal :** un accès simple aux ressources sémantiques (thésaurus, terminologies, vocabulaires, ontologies) est essentiel dans OpenMinTeD pour faciliter leur utilisation par les chaines de traitement de TDM. AgroPortal est un portail de ressources sémantiques – décrites dans des formats standards tels que SKOS ou OWL – pour l'agronomie, les plantes, la nutrition et la biodiversité. Il héberge une centaine de ressources et offre un ensemble de services tels que la recherche, la navigation, l'alignement, et l'annotation de texte. Dans le cadre de VisaTM nous développons un composant d'interconnexion qui permet de décrire les ressources sémantiques d'AgroPortal dans le format de métadonnées OMTD-SHARE et de les importer automatiquement dans la plateforme. L'interconnexion, implémentée sous forme de web service REST, repose sur la technologie partagée de portail d'ontologies développée à Stanford dans le cadre du NCBO (National Center for

[3] On parle également de chaines de traitement ou flux TDM.

[4] ISTEX est financé par les Investissement d'Avenir et repose sur un partenariat entre le CNRS, l'ABES (Agence Bibliographique de l'Enseignement Supérieur), le Consortium COUPERIN et l'Université de Lorraine agissant pour le compte de la CPU

Biomedical Ontology) BioPortal ; ainsi, le composant a été généralisé aux autres portails d'ontologies suivants : NCBO BioPortal (biomédecine, ressources principalement en anglais), SIFR BioPortal (biomédecine, ressources en français), et BiblioPortal (bibliothèques et standards de métadonnées).[5]

L'ensemble de ces interconnexions a permis de dessiner les bases d'une infrastructure ouverte capable de délivrer des services avancés de Text Mining à destination de la recherche scientifique française et par ailleurs de poser les conditions de l'interopérabilité.

[5] NCBO BioPortal : http://bioportal.bioontology.org, SIFR BioPortal: http://bioportal.lirmm.fr, BiblioPortal: https://biblio.ontoportal.org.

# Modélisation de connaissances métier pour la classification de pièces de bois.

Radouan Dahbi[1,2], Vincent Bombardier[1], David Brie[1] et Eric Masson[2]

[1] Laboratoire CRAN, Université de Lorraine,
radouan.dahbi@univ-lorraine.fr, vincent.bombardier@univ-lorraine.fr,
david.brie@univ-lorraine.fr
[2] CRITT Bois, Département Vosges,
88000 Épinal
http://www.crittbois.com

**Résumé** : Cet article présente la démarche de modélisation de connaissances expertes, issues du domaine du bois. L'objectif de l'étude vise la création d'une ontologie métier définissant l'ensemble de primitives nécessaires à l'identification par un système de vision des singularités du bois et du classement qualité correspondant. La modélisation de ces connaissances est basée sur l'application de la méthodologie NIAM/ORM, à partir de la définition en langage naturel par les experts métiers, des singularités influant sur la classification finale. L'ontologie obtenue, validée par les experts du domaine, doit servir à la conception et au paramétrage d'un classificateur hiérarchique pour le tri qualité de pièces de bois en chêne massif.

**Mots-clés** : Modélisation de connaissances, ontologie métier, méthode NIAM/ORM, langage naturel, reconnaissance de formes.

## 1 Introduction

Le bois est un matériau dont le rendu naturel est apprécié dans de nombreuses applications (parquets, ameublement, etc.). Ce rendu se fait majoritairement au travers d'une finition (vernis, peinture, etc.), dont la qualité est impactée par l'hétérogénéité du bois.

Les travaux de thèse CIFRE[1] présentés dans cet article s'effectuent dans le cadre d'une collaboration université – entreprise entre le CRAN et le CRITT Bois[2]. Ils concernent la conception d'une méthode de classification hiérarchique de pièces de bois en chêne avant finition, en fonction d'une analyse de l'aspect visuel et physico-chimique de leur surface. La classification s'effectue à partir de données d'entrée provenant de différents capteurs (caméra linéaire couleur intelligente et imageur hyperspectral proche infrarouge) et vise à prendre en compte la connaissance métier exprimée par les experts du domaine du bois sur les classes de sortie (classe de qualité métier).

## 2 Méthode de modélisation NIAM/ORM

Dans cet article, nous nous intéressons à la modélisation de la connaissance experte liée aux singularités du bois. Cette connaissance est exprimée, par l'expert, en langage naturel en des termes « métier » qui sont quelquefois imprécis et donc peuvent s'avérer subjectifs. Nous entreprenons une formalisation objective de cette connaissance métier en utilisant la méthode NIAM (Natural language Information Analysis Method) (Habrias, 1988) et son formalisme

[1] Contrat CIFRE du 01/12/2017 au 30/11/2020.
[2] Projet OPTIFIN (2015-2019) financé par l'ANR (convention ANR-15-CE10-0007).

ORM (Object Role Modelling) (Halpin, 1998). La littérature propose plusieurs autres méthodes pour formaliser de la connaissance en utilisant le langage naturel comme, Conceptual Graphs (CG) (Dibie-Barthélemy et al., 2006), Object Conceptual Prototyping Language (OCPL) (James et Shipley, 2000). Notre choix s'est porté sur la méthode NIAM/ORM à cause de sa méthodologie déterminant un modèle conceptuel unique, cohérent et contenant toutes les règles nécessaires à une bonne représentation de la réalité. De plus, elle n'est pas spécifique à un domaine et son formalisme ORM fait preuve de simplicité d'utilisation et de mise en œuvre au travers de l'outil VisioModeler qui fournit une modélisation compréhensible pour les non-initiés et dont le module de verbalisation permet de valider de la connaissance émise.

La méthode NIAM/ORM repose sur l'acquisition des connaissances à partir d'un énoncé en langage naturel et leur modélisation sous forme d'un modèle conceptuel. Afin de s'assurer de la cohérence de ce modèle, elle permet l'ajout de contraintes (unicité, totalité, …) sur les objets et leurs relations en posant à l'expert des questions précises. Le modèle de connaissances NIAM/ORM ainsi obtenu peut être, finalement, soumis à validation par l'expert après transcription en langage naturel. Pour plus d'informations sur la méthode, le lecteur peut se référer à (Blaise, 2000).

## 3 Application à la modélisation des singularités du bois

La méthode NIAM/ORM pour la modélisation des connaissances du bois, a déjà été appliquée par (Bombardier et al., 2007) et (Almecija et al., 2012) afin de modéliser la connaissance experte définissant les singularités sur des résineux dans le cadre de la création d'un système de vision pour du contrôle qualité. L'objectif a été dans les deux cas d'aider à la définition des primitives de décision pour l'identification des classes de qualité.

Selon ces mêmes approches, nous visons à créer notre modèle de connaissances NIAM/ORM, en demandant à l'expert du domaine du bois de nous fournir une liste de définitions, en langage naturel, des singularités du bois présentes sur l'essence considérée (dans notre cas le chêne), tout en se référant aux normes européennes relatives à ces singularités. L'objectif de ce modèle est d'aboutir à l'élaboration d'une ontologie métier définissant les caractéristiques à prendre en compte pour identifier chacune des singularités détectées sur les pièces de bois.

Par exemple, l'expert bois décrit les singularités du bois de type « nœud », « picot » et « patte de chat » de la manière suivante :

« *Nœud est une partie de branche englobée dans le bois. Elle apparaît sur le bois scié sous une forme plutôt ronde, qui peut être aussi ovale ou allongée et se caractérise par un diamètre. Sa couleur peut être semblable au bois naturel ou bien partiellement ou totalement noire.* »

« *Picot est un nœud rond ou ovale, sain ou noir, ayant un diamètre maximal de 5 mm*»

« *Patte de chat est un ensemble de picots très rapprochés les uns des autres.* »

L'analyse de plusieurs de ces définitions, permet de faire apparaitre qu'une singularité nommée « nœud » peut être caractérisée par une forme qui peut prendre une et une seule valeur (contrainte d'unicité) parmi « Ronde », « Ovale » ou « Allongée ». Cette singularité est également caractérisée par une couleur et un diamètre. Elle fait également ressortir que les singularités « picots » et « pattes de chat » peuvent être directement déduites de la singularité nœud. Ainsi une hiérarchie apparait, montrant que l'on doit d'abord identifier les « nœuds » puis les « picots » et enfin les « pattes de chat » La figure 1 représente le modèle de connaissances NIAM/ORM correspondant.

La validation de ce modèle de connaissances est effectuée, après transcription en langage naturel, par l'expert qui a émis ces connaissances. Un ensemble de cycles « auteurs – lecteurs » a permis de préciser les notions de contraintes.

La classification des pièces de bois, qui est l'objectif principale de la thèse, s'effectuera par un système de vision. La mise en correspondance des connaissances du bois modélisées avec des connaissances issues du domaine de la vision est donc une nécessité. Cela revient à recenser et modéliser les connaissances de l'expert vision et à les lier aux connaissances de l'expert bois,

de manière à déterminer les paramètres utiles à la quantification des caractéristiques des singularités du bois à identifier. Pour la modélisation des connaissances de l'expert vision, basée aussi sur la méthode NIAM/ORM, nous nous inspirerons de ce qui a été fait dans les travaux de (Bombardier et al., 2007).

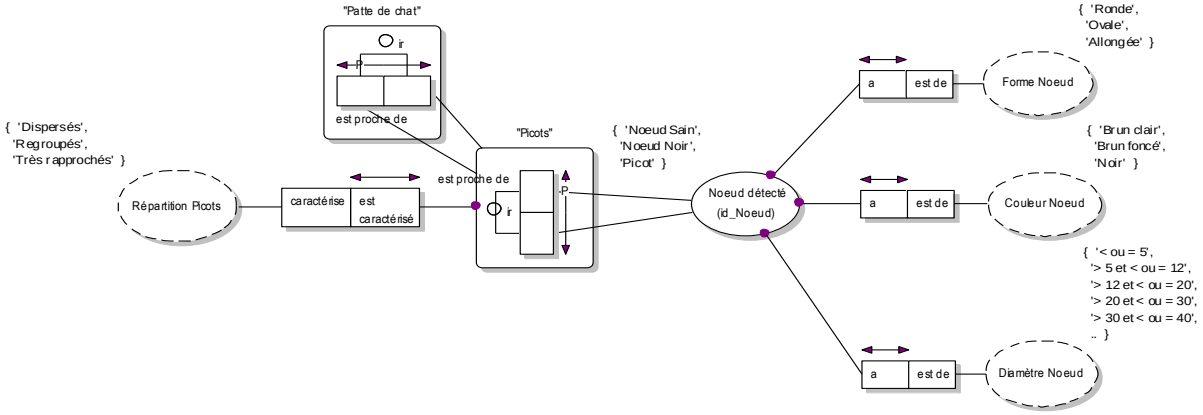



***FIGURE 1*** *– Modèle de singularités de type nœud, picot et patte de chat.*

## 4 Conclusion et perspectives

La modélisation des connaissances liées aux singularités influant sur la classification finale des pièces de bois est en cours et conduira à l'élaboration d'une ontologie métier. L'utilisation de la méthode NIAM/ORM en correspondance avec le formalisme OWL (Ontology Web Language) (Hodrob et Jarrar, 2010) est envisageable pour développer et manipuler cette ontologie dont l'exploitation devrait permettre de définir de manière objective chacune des singularités à identifier, sur les pièces de bois en chêne, par le système de vision. Elle permet de faire apparaitre une hiérarchie entre les singularités et doit ainsi guider la conception du classificateur final. Elle contribue également à capitaliser le savoir-faire de l'industriel.

La formalisation de ces connaissances expertes par la méthode NIAM/ORM montre cependant ses limites dès lors qu'il faut exprimer des contraintes dites « procédurales ». Ces contraintes font partie de la modélisation des classes de sortie, qui représente un travail à venir et dont l'objectif est d'essayer de modéliser le raisonnement des experts bois pour le classement qualité des pièces de bois. Une autre perspective est de déduire « automatiquement » de cette modélisation une arborescence servant comme base à la méthode de classification que nous souhaitons hiérarchique pour des raisons de rapidité de calcul et d'interprétabilité.

## Références

ALMECIJA B., BOMBARDIER V., CHARPENTIER P. (2012). Modeling Quality knowledge to design log sorting system by X rays tomography. *14th IFAC Symposium on Information Control Problems in Manufacturing*, p. 1190-1195. Bucarest, Roumanie.

BLAISE J.C. (2000). *Apport d'une modélisation de l'information normative à l'intégration des règles de sécurité des machines en conception*. Thèse d'université, Université Henri Poincaré, Nancy 1.

BOMBARDIER V., MAZAUD C., LHOSTE P., VOGRIG R. (2007). Contribution of fuzzy reasoning method to knowledge integration in a defect recognition system. *Computers in Industry*, vol. 58, p. 355-366.

DIBIE-BARTHÉLEMY J., HAEMMERLÉ O., SALVAT E. (2006). A semantic validation of conceptual graphs. *Knowledge-Based Systems*, vol. 19, p. 498-510.

HABRIAS H. (1988). *Le modèle relationnel binaire : méthode I.A. (NIAM)*. Éditions Eyrolles.

HALPIN T.A. (1998). Object-Role Modeling (ORM/NIAM). *Handbook on Architectures of Inf. Systems*.

HODROB R., JARRAR M. (2010). ORM to OWL 2 DL Mapping. *International Conference on Intelligent Semantic Web: Applications and Services*, p. 131-137. ACM.

JAMES A.E., SHIPLEY S.D.E. (2000). The development of OCPL, object conceptual prototyping language. *Information and Software Technology*, vol. 42, p. 1045-1056.

# Vers une modélisation des tâches pour l'assistance à la navigation et la reconception de sites Web

Benoît Encelle[1], Karim Sehaba[2]

[1] Université de Lyon,
CNRS Université Lyon 1, LIRIS,
UMR5205, F-69622, France
benoit.encelle@liris.cnrs.fr
Université de Lyon, CNRS
Université Lyon 2
LIRIS, UMR5205, F-69676, France
karim.sehaba@liris.cnrs.fr

**Résumé** : Ce travail s'inscrit dans le cadre des systèmes d'assistance à la navigation Web et à la reconception d'applications Web. Pour l'assistance à la navigation Web par exemple, dans la majorité des systèmes, les aides fournies sont prédéfinies durant la phase de conception et correspondent ainsi aux usages *prévus*. Ces assistances ne tiennent par conséquent souvent pas compte de l'évolution des besoins et des spécificités des utilisateurs. Pour remédier à cette limite, notre objectif est de baser l'assistance sur les usages réels/observés. Pour représenter ces usages dans un formalisme directement manipulable par les utilisateurs et opérable par un système d'assistance, cet article expose une approche pour générer des modèles de tâche *haut niveau* issus des domaines de l'IHM/du génie logiciel à partir de traces « bas niveau » représentant ces usages.

**Mots-clés** : assistance, Web, traces d'interaction, modélisation de la tâche, automates à états finis, ConcurTaskTrees (CTT).

## 1 Introduction

Ce travail s'inscrit dans le cadre de l'assistance à la navigation sur le Web et à la reconception d'applications Web. Il s'agit d'être en capacité d'assister non seulement l'utilisateur dans la réalisation d'une tâche de navigation mais aussi le concepteur dans l'adaptation, la reconception de son application Web. Dans la majorité des systèmes d'assistance existants - pour la navigation Web par exemple - les aides fournies, et en générale les connaissances d'assistance, sont prédéfinies durant la phase de conception et correspondent aux usages prévus. Néanmoins, il est souvent difficile d'appréhender a priori pour un système donné, et une application Web plus particulièrement, tout le spectre des usages réels : différentes populations d'utilisateurs, besoins des utilisateurs en constante évolution, différentes conditions d'utilisation, etc. Si des outils et méthodologies existent pour prévoir certains usages (études/recueils des besoins, prototypages rapides et évaluations en situation écologique), ceux-ci resteront des usages prévus et peuvent dans certains cas ne pas couvrir ou correspondre à l'ensemble des usages réels. Cela peut être dû à plusieurs difficultés liées à : l'analyse de l'ensemble des contextes d'usages, la représentativité de l'échantillon observé, l'obtention de conditions réellement écologiques lors d'évaluations, etc. Ainsi, la conception d'un système d'assistance disposant d'une représentation complète des besoins des futurs utilisateurs et de leurs évolutions est par définition très difficile, voire impossible.

Pour remédier à ces difficultés, nous proposons de baser l'assistance sur les usages réels/observés. Plus précisément, notre approche vise à la génération de modèles de tâche à

partir de ces usages réels. En entrée du processus de génération de modèles, nous partons donc des usages réels, représentés à l'aide de traces d'interaction. Une trace représente les actions d'un utilisateur donné sur une application Web. En sortie, un modèle de tâche représente les différentes possibilités de réalisation d'une tâche donnée. De manière plus précise, un modèle de tâche est une représentation graphique ou textuelle issue d'un processus d'analyse, dont l'objectif est de décrire de manière logique les activités à mener par un utilisateur, voire éventuellement plusieurs, sur l'interface d'un système pour atteindre un but précis. Les modèles de tâche obtenus par le processus proposé seront ensuite employés dans un système d'assistance pour guider les utilisateurs dans l'accomplissement de leurs tâches et les concepteurs dans l'analyse de leurs applications et dans d'éventuelles reconceptions de celles-ci.

Nous avons identifié deux propriétés principales devant être supportées par les modèles de tâche pour que ceux-ci puissent être employés à des fins d'assistance. Il s'agit de : 1/ l'intelligibilité du modèle par l'utilisateur, et 2/ l'expressivité du modèle de tâche et sa capacité d'exploitation par un système informatique, plus particulièrement un système d'assistance.

Dans la littérature, plusieurs métamodèles et notations ont été proposés pour représenter des modèles de tâche. Par conséquent, nous avons confronté ensuite ces métamodèles aux caractéristiques précédemment identifiées pour déterminer ceux ou celui qui nous semble être le plus adapté à notre objectif d'assistance. Cette étude nous a permis de choisir le métamodèle ConcurTaskTrees (CTT) (Paternò, 2003). Enfin, nous avons développé par la suite un processus de génération de modèles de tâche, représentant les usages réels, en se basant sur des traces d'interaction.

Notre travail développe les trois contributions suivantes :

1 La spécification des caractéristiques des métamodèles pour des objectifs d'assistance ;
2 La confrontation des métamodèles existants au regard des caractéristiques identifiées (CTT a été choisi) ;
3 Un processus de génération de modèles de tâche CTT à partir de traces. Les algorithmes développés permettent pour l'instant l'identification des opérateurs CTT d'activation et d'indépendances.

Dans nos travaux futurs, nous souhaitons développer d'autres algorithmes de conversion dans l'optique de couvrir tous les opérateurs CTT. Ensuite, nous envisageons premièrement d'évaluer ces algorithmes à l'aide de données de navigation, puis d'évaluer en situation écologique deux systèmes d'assistance basés sur notre approche : un pour l'aide à la navigation et l'autre pour l'aide à la reconception de site.

## Références

Paternò, F. (2003). The Handbook of Task Analysis for Human-Computer Interaction. In (483-503). Taylor & Francis.

# Jeu de données SemBib: représentation sémantique des données bibliographiques de Télécom ParisTech

Jean-Claude Moissinac*

*LTCI, Télécom ParisTech, Université Paris-Saclay, 75013, Paris, France
jean-claude moissinac@telecom-paristech.fr,
https://moissinac.wp.imt.fr/

**Résumé.** Nous allons présenter ici le jeu de données SemBib, représentation sémantique des données bibliographiques de Télécom ParisTech. Ce travail est mené dans le cadre du projet SemBib, au sein de Telecom ParisTech.

## 1 Introduction

Nous allons présenter ici le jeu de données SemBib, représentation sémantique des données bibliographiques de Télécom ParisTech. Ce jeu de données est en libre accès.

SemBib vise à constituer et exploiter une base de connaissances sur nos publications scientifiques avec plusieurs objectifs : pouvoir mieux exploiter notre production scientifique ; améliorer la qualité de notre base bibliographique ; pouvoir nous interconnecter avec d'autres bases ; disposer d'un jeu de données riche pour nos travaux d'étudiants, mais, aussi, nous l'espérons, pour d'autres expérimentations sur les données bibliographiques. SemBib recense actuellement des metadonnées pour 11311 publications et dispose des textes intégraux de 4939 publications (ces derniers ne sont pas encore accessibles librement).

## 2 Contexte

Un grand intérêt a été porté ces dernières années pour la création d'outils facilitant l'exploration d'un nombre en croissance rapide de publications scientifiques. Nous avons connaissance de multiples initiatives pour constituer des outils autour de bases bibliographiques : CrossRef, DBLP, HAL, Google ScholarMicrosoft Academic Graph...

Les cas de DBLP et HAL illustrent l'intérêt que nous portons à la constitution d'une base propre. Pour Télécom ParisTech, DBLP ne connait qu'environ 300 publications et HAL environ 3000, alors que nous en avons plus de 11000.

Une analyse de nos publications nous a montré que 53 formes ont été utilisées au cours du temps par nos chercheurs pour indiquer leur affiliation. Le rapprochement entre ces désignations n'est pas chose facile pour un organisme externe. Nous pouvons aisément faire ces rapprochements. Ce type d'observation peut être décliné pour les noms d'auteurs et les canaux de publication. Notre approche de consolidation d'une base interne nous parait être une étape nécessaire pour être bien représenté dans des bases externes comme HAL.

TAB. 1 – *Exemple de représentation de publication (en turtle, préfixe non déclarés pour simplicité)*

```
<http ://givingsense.eu/sembib/onto/tpt/biblio/14557>
        rdf :type               ns0 :ResearchPaper ;
        ns2 :firstAuthor        <http ://givingsense.eu/sembib/onto/persons/Angelini> ;
        ns2 :state              "published" ;
        ns3 :entrytype          "article" ;
        ns3 :fromDpt            <http ://givingsense.eu/sembib/onto/tpt/TSI> ;
        ns3 :fromGroup          <http ://givingsense.eu/sembib/onto/tpt/TII> ;
        ns3 :ref                "YH :TMI-14" ;
        ns4 :creator            <http ://givingsense.eu/sembib/onto/persons/Hoffman> ,
                                <http ://givingsense.eu/sembib/onto/persons/Barr> ,
                                <http ://givingsense.eu/sembib/onto/persons/Angelini> ;
        ns4 :language           "en" ;
        ns4 :title              "Adaptive quantification... kov measure field mode" ;
        ns1 :publicationDate    "2014"|http ://www.w3.org/2001/XMLSchema#integer> ;
        ns2 :venue              <http ://givingsense.eu/sembib/onto/channels/WIMS_14> ;
        ns6 :Abstract           "The extent of pulmonary ... random fields (MRFs)." .
```

# 3 Choix de représentation

Nous allons ici présenter le modèle de représentation retenu par SemBib.

Au cœur du projet, nous avons un graphe -au sens de la représentation RDF- de représentation de nos publications. Chaque publication est représentée par une entité désignée par une URI à laquelle sont associées un ensemble de propriétés (prédicats selon RDF). Essentiellement, les propriétés utilisées dans ce graphe sont les propriétés intrinsèques d'une publication : titre, auteurs, année de publication, résumé éventuel.

Certaines propriétés ont une valeur littérale, comme le titre, d'autres ont pour valeur une URI vers une entité décrite dans un autre graphe. Ce graphe fait notamment référence au graphe des auteurs et affiliations ainsi qu'au graphe des canaux de publication. La table 1 donne un exemple de représentation.

Nous avons choisi d'exploiter des vocabulaires bien identifiés pour ce type de données. Le Dublin Core est un point de départ. Au niveau bibliographique, nous avons trouvé que la famille d'ontologies SPAR (Shotton et al. (2009)) constituait un ensemble solide sur lequel construire ; nous avons notamment appuyé notre choix sur l'analyse de Ruiz-Iniesta et Corcho (2014). Pour cette raison, à l'avenir, les évolutions successives de nos graphes vont intégrer de plus en plus de concepts définis par les ontologies SPAR.

# 4 Données initiales et données SemBib

Actuellement, nous travaillons sur 11311 publications référencées [1] dans notre base bibliographique depuis 1969. Au-delà des meta-données -titre, auteurs,lieu de publication, année-, la

1. en novembre 2017

base fournit seulement 1313 URL d'accès à la publication proprement dite. Nous avons mis en place des automatismes en vue de collecter l'ensemble de nos publications. A ce jour, environ 5000 textes intégraux ont été récupérés.

L'exploitation de ces données initiales a permis de construire

— un graphe de représentation des publications qui comporte actuellement 263147 triplets [2],
— un graphe de description des personnes et organisations impliquées dans nos publications qui comporte actuellement 50411 triplets[2],
— un graphe de description des canaux de publications que nous avons utilisé ; il comporte actuellement 6599 triplets[2].
— des graphes génériques de concepts à relier aux articles, aux auteurs et aux canaux de publication.

Ces graphes -enrichis au fil des travaux- sont accessibles sur un point d'accès SPARQL [3] et un dump sur github [4] en est assuré périodiquement.

Notre approche a ainsi comme conséquences d'améliorer la qualité des données de notre base bibliographique interne, de nous permettre de disposer d'informations qui n'ont pas leur place dans les bases externes (groupes de recherche ou sein des départements, projets, ...), de nous interconnecter avec d'autres bases sur les principes du LOD -Linked Open Data-, d'être une source de qualité pour alimenter des bases génériques comme HAL, de bénéficier d'une meilleure indexation de nos publications par les moteurs de recherche.

# 5 Conclusion

Nous avons présenté un jeu de données conçu comme base de travail pour des méthodes de représentation sémantique de données bibliographiques dans le domaine des technologies de l'information. Nous pensons que nos choix pour la représentation ouvrent un large champs pour l'exploration, l'exploitation et l'interconnexion de ces données.

# Références

Ruiz-Iniesta, A. et Ó. Corcho (2014). A review of ontologies for describing scholarly and scientific documents. In A. G. Castro, C. Lange, P. W. Lord, et R. Stevens (Eds.), *Proceedings of the 4th Workshop on Semantic Publishing co-located with the 11th Extended Semantic Web Conference (ESWC 2014), Anissaras, Greece, May 25th, 2014.*, Volume 1155 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Shotton, D., K. Portwin, G. Klyne, et A. Miles (2009). Adventures in semantic publishing : exemplar semantic enhancements of a research article. *PLoS Comput Biol 5*(4), e1000361.

2. au 8/4/2018
3. `http://givingsense.eu/sembib/sparql/` ce point d'accès est utilise actuellement ARC2, qui n'implémente que partiellement SPARQL
4. `https://github.com/moissinac/sembib-graphs`

# Intégration d'ontologies médicales : améloriation par association des maladies humaines à leurs plus pertinents signes caratéristiques

Adama Sow, Abdoulaye Guissé, Oumar Niang

LABORATOIRE TRAITEMENT DE L'INFORMATION ET SYSTÈMES INTELLIGENTS (LTISI)
Département du Génie Informatique et Télécommunications
Ecole Polytechnique Thiès (EPT) , THIES, Sénégal
asow@ept.sn, aguisse@ept.sn, oniang@ept.sn

**Résumé** : Guérir un patient malade nécessite un diagnostic médical avant de proposer un traitement approprié. Avec l'explosion des connaissances médicales, nous nous intéressons à leur exploitation pour aider le médecin à collecter des informations et prendre des décisions lors du processus de diagnostic médical. Le présent article propose une ontologie à partir d'une intégration de plusieurs ontologies et terminologies médicales existantes et ouvertes. Nous constituons une nouvelle ontologie fédératrice couvrant toutes les maladies humaines en incluant des liens avec leurs signes caractéristiques pertinents. Pour prendre en compte cette association, l'ontologie produite est alimentée continuellement par apprentissage des signes à partir d'une base de cas réels de diagnostics cliniques confirmés et validés.

**Mots-clés** : diagnostic médical, ontologies médicales, intégration d'ontologies, web de données, systèmes e-santé

## 1 Introduction

Le diagnostic médical, tel que décrit dans le livre de Balogh *et al.* (2015) est une activité cognitive centrée sur le patient dont la compétence quintessentielle appartient au médecin. C'est un procédé qui consiste en une collecte continue des informations médicales qu'effectue le médecin avant de les intégrer et de les interpréter pour la gestion des problèmes de santé de son patient. Cette première étape de collecte est aussi capitale que complexe pour le médecin surtout lorsque cela nécessite de recourir rapidement, en un temps réduit, à des masses de connaissances médicales qui ne cessent d'exploser à l'échelle internationale. C'est dans l'optique d'assister les médecins dans l'exploitation de ces connaissances, que se situe notre recherche. Nous mettons ici le focus sur le nombre important d'ontologies médicales, et l'objectif est d'avoir une ontologie centre qui réportorie toutes les informations pertinentes dans l'élaboration d'un diagnostic médical.

Les ontologies médicales ont été conçues (Hoehndorf *et al.* (2015); Anbarasi *et al.* (2013)) pour mettre en place des vocabulaires médicaux communs reposant sur des concepts partagés qui facilitent l'interopérabilité des documents entre les acteurs du domaine et surtout l'élaboration des connaissances. Nous nous intéressons ici aux ontologies médicales des maladies humaines. La liste est longue et chaque ontologie présente ses propres spécificités. Mais globalement toutes les maladies sont couvertes et renvoient chacune à un concept regroupant ses divers termes nominatifs et leurs synonymes, ses différentes définitions et axiomes textuels et ses signes caractéristiques. Ces derniers indiquent entre autres des signes cliniques et des symptômes (Cox *et al.* (2014)), mais aussi éventuellement l'agent en cause de la maladie, le mode de transmission, et la localisation dans l'anatomie humaine.

Le présent article porte sur une fédération de diverses ontologies. En effet, dans les ontologies existantes nous trouvons d'une part des ontologies de maladies associées à des signes généraux dont l'exhaustivité est à éclaircir, et d'autre part des ontologies qui conceptualisent tous les signes susceptibles d'être identifiés chez un malade mais aucun lien avec les maladies concernées n'est identifié. Nous nous intéressons alors à une intégration de ces ontologies afin de lier chaque maladie à ses plus pertinents signes. Pour prendre en compte cette association, l'ontologie produite est alimentée continuellement par apprentissage des signes à partir d'une base de cas de diagnostics cliniques.

## 2 Méthodologie de fédération

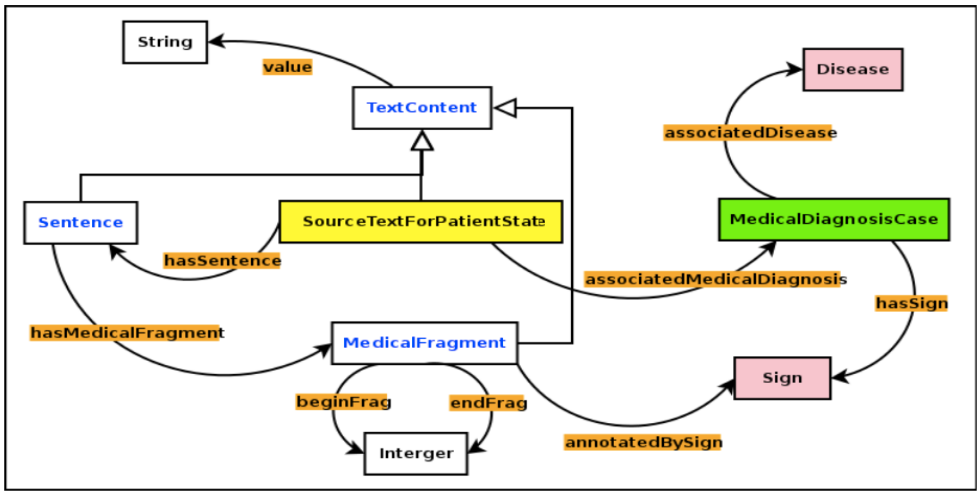



FIGURE *1 – Structure de données d'ensemble*

La constitution d'une ontologie de maladies et de signes consiste à une fédération (Figure 1) d'un ensemble d'ontologies autour d'une structure unifiant toutes les maladies humaines ainsi que leurs signes caractéristiques. Les maladies correspondent aux diagnostics possibles. Les signes sont ceux susceptibles d'être identifiés sur un patient afin de conclure sur un diagnostic précis qui lui peut renvoyer à une ou plusieurs maladies.

Les maladies sont organisées de façon hiérarchisée ; elles et leurs formes dérivées sont regroupées par catégories, qui peuvent elles-mêmes être composées de sous-catégories de maladies. Les maladies sont lexicalisées afin d'avoir pour chaque maladie l'ensemble des termes nominatifs les plus connus et leurs synonymes. Pour chaque maladie, il sera important de conserver les définitions afin de contrôler la sémantique la mieux partagée. La plupart des signes connus de chaque maladie sont listés formellement à partir de ceux disponibles dans les ontologies médicales cibles.

Nous analysons ici des ontologies médicales mises à la disposition du public via la plateforme BioPortal. Notre choix s'est porté sur la DOID, la MESH, la SNOMED comme ontologies de maladies, ainsi que la SYMP, et la CSSO comme ontologies de signes. L'ontologie est chargée (voire tableau 1 de la Figure 2) par interrogation de ces différentes ressources ontologiques cibles avec le langage de requêtes SPARQL à partir de la plateforme BioPortal. Nous sélectionnons ainsi toutes les maladies qui constituent les feuilles des classes à partir de la DOID, ainsi que leurs définitions à partir de MESH ; toutes les catégories de maladies à partir de la DOID où nous sélectionnons leur nom, leur description, et leurs catégories mères ; tous les termes nominatifs synonymes des maladies à partir de la DOID, mais surtout à partir de MESH, soient le label préférentiel, ainsi que les labels alternatifs pour chaque maladie ; tous les signes caractéristiques de base pour chaque maladie à partir des descriptions semi-formalisées de la DOID ; et enfin tous les termes nominatifs synonymes des signes : les labels préférentiels sont extraits de SYMP, les labels alternatifs sont quant à eux extraits des ontologies CSSO, et SNOMED.

Après avoir constitué notre fédération d'ontologies, nous enrichissons l'ontologie produite à partir de l'analyse de rapports médicaux de cas de diagnostics ayant déjà été validés par des médecins. Le tableau 2 (voire Figure 2) montre les symptômes et les signes cliniques trouvés sur une dizaine de cas réels de patients ayant déjà été diagnostiqués. Les exemples choisis portent sur des maladies tropicales que nous trouvons au Sénégal. Nous pouvons alors remarquer que pour chaque maladie, il y a un nombre précis de symptômes généraux indiqués par notre ontologie de maladies mais la totalité d'entre eux ne sont pas présents chez les

patients. De nouveaux symptômes non répertoriés dans l'ontologie font leur apparition ainsi que les signes cliniques dont les valeurs sont spécifiques à chaque patient.

| Types Eléments | Objet ontologique | Ontologies d'origines | Nbr. d'individus |
|---|---|---|---|
| **Diagnostics** | | | |
| Maladies | Disease Class | DOID | 6442 |
| Catégories | SetOfDiseases Class | DOID | 3947 |
| Termes Synonymes | AnnotationProperty (prefLabel, altLabel, hiddenLabel) | DOID, MESH | 27586 |
| **Signes** | | | |
| Symptômes et Signes Cliniques | Symptom Class et CinicalSign Class - subClassOf Sign Class | SYMP | 942 |
| Autres signes | PhysicalAgent Class, ChimicalAgent Class, TopographicalLocate Class, MedicalProcedure Class : subClassOf Sign Class | DOID, SNOMED | 6020 |
| Termes Synonymes | AnnotationProperty (prefLabel, altLabel) | CSSO, SNOMED | 1346 |

TABLE *1 – Description de l'état actuel de notre ontologie de maladies et de signes*

| Maladie | Symptômes indiqués par l'ontologie | Symptômes de l'ontologie présents sur le cas | Symptômes nouveaux | Signes cliniques |
|---|---|---|---|---|
| Hepatitis A | 9 | 7 | 7 | 9 |
| Cholera | 5 | 3 | 10 | 2 |
| Rougeaole | 6 | 4 | 11 | 3 |
| Dengue | 10 | 5 | 11 | 20 |
| Tétanos | 4 | 3 | 12 | 4 |
| Paludisme | 6 | 4 | 8 | 24 |
| Syphilis | 5 | 2 | 12 | 14 |
| Chikungunya | 9 | 5 | 7 | 29 |
| FièvreTyphoide | 8 | 5 | 8 | 3 |
| Meningite | 9 | 4 | 5 | 7 |

TABLE *2 – Nombre de Simptômes et de Signes cliniques trouvés sur des cas réels*

FIGURE *2 – L'ontologie résultantes en chiffres*

## 3 Conclusion

Dans cet article, la problématique porte sur la mise en place d'un système d'exploitation des ressources ontologies ouvertes et partagées. Il est question ici de constitution d'une ontologie centrale fédérant un ensemble d'ontologies et terminologies médicales cibles, qui répondent au besoin en informations afin de faciliter la tâche du médecin dans l'identification des diagnostics potentiels, parmi lesquels il aura la latitude de choisir ou de valider le plus fiable en connaissance de cause. Ce type de système ne se substitue donc nullement au médecin. Nous avons donc proposé une méthodologie d'intégration autour d'une structure de graphe RDF facilitant la récupération des maladies humaines et de leurs plus pertinents signes caractéristiques, à partir des ontologies cibles et d'une analyse de cas réels de diagnostics confirmés par des médecins. Au final, nous disposons d'une ontologie de maladies et de signes qui sert de base de connaissances pour l'aide au diagnostic médical.

## Références

ANBARASI M., NAVEEN P., SELVAGANAPATHI S. & NOWSATH M. (2013). Ontology based medical diagnosis decision support system. In *Actes de International Journal of Engineering Research and Technology (Traitement automatique des langues naturelles)*.

BALOGH E. P., MILLER B. T. & BALL J. R. (2015). Improving diagnosis in health care. In *Actes de National Academies of Sciences, Engineering, and Medicine. The National Academies Press, Washington, DC (Traitement automatique des langues naturelles)*.

COX A. P., RAY P. L., JENSEN M. & DIEHL A. D. (2014). Defining 'sign' and 'symptom'. In *Actes de IWOOD Workshop, In ICBO, Houston, TX, USA, October 6-7, 2014 (Traitement automatique des langues naturelles)*, p. 101–110.

HOEHNDORF R., SCHOFIELD P. & GKOUTOS G. (2015). The role of ontologies in biological and biomedical research : a functional perspective. In *Actes de Briefings in Bioinformatics Journal, 2015 (Traitement automatique des langues naturelles)*.

# Vers une recommandation personnalisée de ressources pour l'apprentissage en ligne

Sarra Bouzayane[1], Inès Saad[1,2]

[1] LABORATOIRE MIS, Université de Picardie Jules verne, Amiens, France
{sarra.bouzayane,ines.saad}@u-picardie.fr

[2] Ecole Supérieure de Commerce, Amiens, France

**Résumé** : Ce travail propose un système dédié aux apprenants des MOOC (Massive Open Online Courses) pour recommander à chacun d'entre eux une liste personnalisée de ressources pédagogiques appropriées à son profil afin d'améliorer son processus d'apprentissage médiatisé. Notre système repose sur une approche de prédiction périodique des "Apprenants leaders" et une phase de recommandation basée sur le filtrage basé connaissance. Les ressources recommandées sont uniquement celles déposées par des "Apprenants leaders".

**Mots-clés** : Systèmes de recommandation, Transfert de connaissances, MOOCs.

## 1 Introduction

Ce travail s'intéresse aux systèmes de recommandation dans le domaine d'apprentissage médiatisé, en particulier les MOOCs (*Massive Open Online Courses*). Ce sont des formations en ligne et gratuites accessibles par un nombre massif d'apprenants et animées par une équipe pédagogique de taille réduite ce qui rend difficile le processus d'accompagnement. Ainsi, pour répondre à leurs questions, les apprenants ont recours généralement aux informations échangées sur le forum du MOOC dont l'exactitude n'est pas toujours garantie.

Pour ce faire, nous recommandons à ces apprenants une liste personnalisée des ressources pédagogiques. Afin de garantir la qualité des ressources, nous recommandons uniquement celles déposées par des "Apprenants leaders". Cette catégorie d'apprenants est prédite d'une manière hebdomadaire en appliquant notre méthode de classification incrémentale MAI2P (Bouzayane & Saad, 2017) basée sur l'approche DRSA (*Dominance based Rough Set Approach*) (Greco *et al.*, 2001). Le système repose sur deux modélisations : la modélisation des profils apprenants en fonction de leurs préférences et la modélisation des ressources pédagogiques selon leurs taux de pertinence liés aux profils des apprenants les ayant déposées.

## 2 Un système de recommandation pour les apprenants des MOOCs

Le processus de recommandation adopté par le système *KTI-MOOC* (*Recommender system for the Knowledge Transfer Improvement within a MOOC*) repose sur la technique de filtrage basé-connaissance. Il est composé de trois étapes : (1) la modélisation des ressources, (2) la modélisation du profil apprenant, et (3) le calcul de taux de correspondance entre eux.

### 2.1 Modélisation d'une ressource

Dans le système KTI-MOOC une ressource pédagogique est modélisée par trois critères :

— Thème : un MOOC diffusé doit porter sur un seul cours décomposé en chapitres. Cette structuration nous a permis d'identifier un ensemble de thèmes pouvant caractériser une ressource pédagogique. Lorsque un apprenant souhaite déposer une ressource, le système doit lui fournir une liste des thèmes permettant de la caractériser.

— Type : lors de dépôt d'une ressource, l'apprenant doit aussi l'indexer par son type en utilisant une liste déroulante fournie par le système. Une ressource peut être visuelle(graphiques, flèches, etc.), orale (vidéos, audio) ou écrite (pdf, word, etc.).

— Pertinence : une ressource est considérée pertinente si elle est déposée par un "Apprenant leader" ayant un état de connaissance acceptable sur ses thèmes. L'état de connaissance d'un apprenant sur chaque thème décrivant la ressource doit être précisé par lui même lors de dépôt de la ressource. Deux états sont définis : 0 si l'apprenant n'a aucune connaissance ou une connaissance moyenne et 1 sinon. La pertinence d'une ressource peut évoluer d'une semaine à une autre en fonction de l'évolution du profil de l'apprenant (catégorie et état de connaissance). La pertinence d'une ressource *i* à un instant *t* est calculée par : (1) le profil de l'apprenant qui l'a soumise à l'instant *t* (cette valeur prend 1 si l'apprenant est leader et 0 sinon) noté $ProfileApp_t$ ; et (2) la moyenne de son état de connaissance à l'instant *t*, noté $ConnApp_{t,k}$, sur chaque thème *k* des *n* thèmes décrivant la ressource, tel que :

$$Pertinence_{(i,t)} = ProfileApp_t * \frac{\sum_{k=1}^{n} ConnApp_{t,k}}{n}$$

Une ressource pédagogique *i* déposée par un apprenant à l'instant *t* sur le site du MOOC est modélisée comme suit :

$$Ressource_{(i,t)} = \{Pertinence_{(i,t)}, Type_i, \sum_{k=1}^{n} Theme_{i,k}\}$$

### 2.2 Modélisation du profil d'un apprenant

Le profil de l'apprenant *j* à l'instant *t* peut être représenté comme suit :

$$Apprenant_{(j,t)} = \{Type_j, \sum_{k=1}^{m} Theme_{j,k}\ \}.$$

tel que $Type_j$ est le type d'apprentissage (visuel, oral ou écrit) préféré par l'apprenant *j* et $Theme_{j,k}$ est le thème *k* des *m* thèmes sur lesquels l'apprenant souhaite avoir de l'aide.

### 2.3 Recommandation

Afin de recommander à un apprenant *j* une ressource *i* qui lui correspond, nous avons appliqué la distance euclidienne comme suite :

$$Rec_{(i,j)} = \sqrt{\sum_{i=1}^{n} (x_i - x_j)^2} \text{ où } Pertinence_{(i,t)} \geqq 0.7$$

tel que $x_i - x_j$= 0 si le thème (ou le type) recherché correspond au thème (ou le type) décrivant la ressource, et $x_i - x_j$= 1 sinon. Uniquement les ressources dont la pertinence est supérieure à 0,7 seront recommandées est classées en fonction de leur taux d'appréciation $Rec_{(i,j)}$.

## 3 Conclusion

Ce papier a proposé une recommandation permettant d'améliorer le processus d'apprentissage en ligne. Le système repose sur deux phases : une phase de prédiction périodique des "Apprenants leaders" et une phase de recommandation. Uniquement les ressources déposées par les apprenants leaders seront recommandées. La pertinence d'une ressource pédagogique évolue en fonction du profil de l'apprenant qui l'a déposée. Nos travaux futurs envisagent améliorer la recommandation en tenant compte du contexte de l'apprenant cible.

## Références

BOUZAYANE S. & SAAD I. (2017). Prediction method based drsa to improve the individual knowledge appropriation in a collaborative learning environment : case of moocs. In *Proceedings of the 50th Hawaii International Conference on System Sciences*, p. 24–33.

GRECO S., MATARAZZO B. & SLOWINSKI R. (2001). Rough sets theory for multicriteria decision analysis. *EJOR*, **129**(1), 1–45.

# Améliorer la recommandation des ressources dans l'apprentissage collaboratif en ligne

Samia Beldjoudi[1], Hassina Seridi[2]

[1] Ecole Supérieure de Technologies Industrielles Annaba, Algérie,
s.beldjoudi@epst-annaba.dz
[1,2] Laboratorie de Géstion Electronique de Documents LabGED, Université Badji Mokhtar Annaba, Algérie
seridi@labged.net

**Résumé** : Récemment, les Linked Open Data (LOD) permettent de créer des liens entre des entités du Web pour lier des informations dans un seul espace de données global. Ce papier démontre comment un contenu structuré accessible via le LOD peut être utilisé pour soutenir la recommandation de ressources éducatives dans les folksonomies. Une des limitations de la recommandation est la surspécialisation du contenu conduisant à l'incapacité de recommander des ressources pertinentes différentes de celles que l'apprenant connaissait auparavant. Pour résoudre ce problème, nous avons proposé de tirer parti de la richesse du graphe DBpedia et Ant Colony Optimization (ACO) pour apprendre le comportement des utilisateurs. L'idée de base est d'explorer itérativement le graphe de données pour produire des recommandations pertinentes et diversifiées. En utilisant ACO, notre système effectue une recherche des chemins appropriés dans le graphe LOD et sélectionne les meilleurs voisins d'un apprenant actif pour fournir des recommandations pertinentes.

**Mots-clés** : Folksonomies, e-learning, recommandation, données ouvertes liées, diversité, optimisation par des colonies de fourmis.

## 1 Introduction

L'objectif principal de notre travail est de savoir comment exploiter l'aspect sémantique de LOD pour améliorer la recommandation de ressources éducatives dans les activités d'étiquetage social. Dans notre processus de recommandation, nous démontrons comment nous pouvons assurer la diversité et la nouveauté dans la recommandation en prenant en compte l'exploration des LOD et l'algorithme ACO. Nous avons donc utilisé la force des données ouvertes liées pour améliorer la recommandation de ressources éducatives dans le système de marquage social en explorant les entités interconnectées dans le cloud LOD.

## 2 Travaux connexes

Dans cette section, un aperçu sur des contributions récentes attachées aux systèmes de recommandation dans le web 2.0 est proposé. Dans (Beldjoudi et al., 2017), les auteurs ont proposé une méthode pour analyser les profils d'utilisateurs en fonction de leurs tags afin de prédire des ressources personnalisées intéressantes et de les recommander. (Karabadji et al., 2018) ont proposé de se concentrer principalement sur la croissance du grand espace de recherche des profils d'utilisateurs et d'utiliser un système de recommandation évolutif multi-objectif basé sur l'optimisation pour extraire un groupe de profils maximisant la similarité

avec le utilisateur actif et la diversité entre ses membres. Dans (Beldjoudi et al., 2018) les auteurs veulent tirer parti du web social et sémantique afin d'améliorer la recommandation de ressources pédagogiques en apprentissage collaboratif.

## 3 Description de l'approche

### 3.1 Exploration de LOD pour assurer la diversité et la nouveauté dans la recommandation

Pour assurer la diversité lors de la recommandation, nous proposons d'émerger les caractéristiques qui ont intéressé l'apprenant quant il a marqué ses ressources. Par exemple, supposant qu'un profil d'un apprenant est composé de ressources (R1, R2, R3 et R4), Ainsi l'intersection des caractéristiques de ces ressources doit être calculée (R1∩ R2 ∩ R3 ∩ R4). Ensuite, pour chaque caractéristique trouvée, nous explorerons le graphe LOD au premier niveau pour extraire d'autres ressources (R5) ayant ces caractéristiques ou ayant un lien direct / indirect avec celles-ci (R6, R7 resp).
Notre approche est basée sur l'exploration itérative du graphe DBpedia. Le but est d'obtenir des recommandations qui devraient non seulement satisfaire l'apprenant mais aussi avoir une diversité et une nouveauté dans le résultat proposé pour créer l'effet de surprise en recommandant des ressources auxquelles l'apprenant ne s'attendait pas au début. L'apprenant évalue les ressources recommandées en temps réel à chaque itération. Le processus s'arrête lorsqu'aucune des ressources recommandées n'a satisfait l'utilisateur.

### 3.2 Utilisation de l'optimisation par des colonies de fourmis pour améliorer la recommandation dans le graphe LOD

La principale contrainte est que dans chaque itération, nous ne choisirons que les ressources du niveau suivant, ceci peut limiter le nombre de ressources à recommander et ignorer complètement l'aspect social dans le processus de recommandation (les voisins de l'utilisateur ne sont pas être considérés). Afin de remédier à ce problème, nous suggérons d'utiliser l'algorithme ACO pour bénéficier de la force de l'aspect communautaire qui caractérise les fourmis. Avec l'utilisation de l'algorithme ACO, nous pouvons recommander plus de ressources à un utilisateur car nous pouvons facilement explorer le graphe LOD sans être limité à un niveau spécifique pendant la recherche, ainsi que l'aspect social de notre approche peut être émergé sans calculer la similitude entre les utilisateurs. Cela se fait comme suit:
Chaque utilisateur est représenté par une fourmi. A chaque fois que l'utilisateur accepte la ressource recommandée par le système, le chemin est marqué par une phéromone. Et donc les ressources fortement acceptées par la majorité des utilisateurs ont plus de phéromone.
Dans ce cas, lorsque nous souhaitons recommander des ressources à un utilisateur actif, le système commence par voir si le chemin de recommandation est marqué par une phéromone supérieure ou égale à un seuil donné. Si c'est le cas, le système recommande directement toutes les ressources de ce chemin à cet utilisateur.

## 4 Les résultats expérimentaux

La base de données exploitée dans notre test est del.icio.us. Elle comprend 1712 annotations impliquant 150 utilisateurs, 543 tags et 744 ressources. Nous avons utilisé DBpedia, l'une des initiatives les plus réussies, basée sur les principes de Linked Open Data.
Afin d'évaluer la qualité de notre système de recommandation, nous avons utilisé les trois mesures: rappel, précision et la métrique F1 dans cinq itérations.
La courbe présentée à la figure 1 montre que la moyenne des trois mesures est bonne dans les cinq itérations.

Pour calculer la diversité et la nouveauté individuelles, nous avons utilisé les formules proposés dans (Zhang et Hurley, 2009) et (Vargas, 2014) respectivement. La figure 2 montre des valeurs prometteuses de diversité et de nouveauté dans les cinq itérations. Cela démontre l'importance de Linked Open Data pour extraire des ressources plus diversifiées et nouvelles lors de la recommandation.

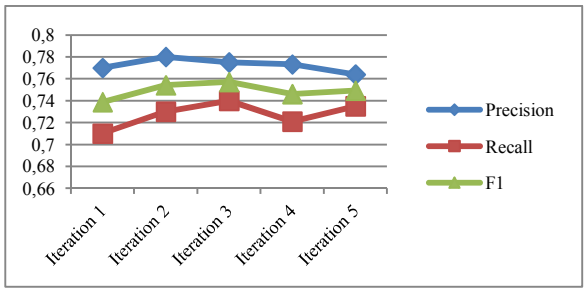



FIGURE 1 – Précision, rappel et F1.

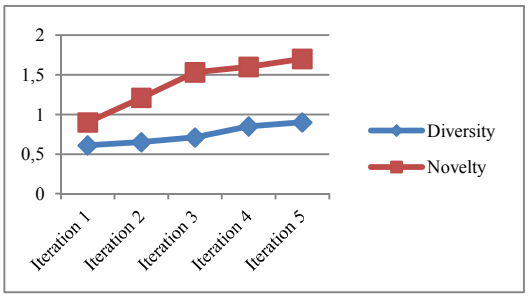



FIGURE 2 – Diversité et nouveauté

## 5 Conclusion

L'objectif de ce travail était de surmonter le problème de la diversité et de la nouveauté dans la recommandation. Les premiers résultats montrent l'utilité d'explorer le graphe LOD pour assurer la diversité lors de la recommandation. Afin de continuer et d'améliorer notre travail, nous visons à utiliser d'autres principes tels que la détection d'événements pour aider à capturer et à analyser le comportement des apprenants lorsque de nouveaux événements surviennent. Cela peut améliorer la recommandation et même le classement des ressources.

## Références


Beldjoudi, S., Seridi, H. & Karabadji, NEI. (2018). Recommendation in Collaborative E-Learning by Using Linked Open Data and Ant Colony Optimization. In: ITS 2018.

Beldjoudi, S., Seridi, H. & Faron-Zucker, C. (2017). Personalizing and Improving Resource Recommendation by Analyzing Users Preferences in Social Tagging Activities. Computing and Informatics 36(1): 223-256 (2017)

Karabadji, NEI., Beldjoudi, S., Seridi, H., Aridhi, S. & Dhifli, W. (2018). Improving memory-based user collaborative filtering with evolutionary multi-objective optimization. Expert Syst. Appl. 98: 153-165.

Vargas, S. (2014). Novelty and diversity enhancement and evaluation in recommender systems and information retrieval. In Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '14, pages 1281–1281, New York, NY, USA.

Zhang, M. and Hurley, N. (2009). Novel item recommendation by user profile partitioning. In 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, volume 1, pages 508–515, Sept.

# Linkky: Extraction de clés de liage par une adaptation de l'analyse relationnelle de concepts

Jérôme David, Jérôme Euzenat, Jérémy Vizzini

Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG, F-38000 Grenoble France
Jerome.David@inria.fr, Jerome.Euzenat@inria.fr, jeremyvizzini@icloud.com

**Résumé** : Les clés de liage permettent de spécifier la manière d'engendrer des liens entre deux sources de données RDF. L'objet de cette démonstration est de présenter Linkky, un prototype implémentant l'extraction de familles de clés de liage dépendantes à l'aide de techniques d'analyse formelle de concept.

**Mots-clés** : Liage de données, clés de liage, RDF, analyse formelle de concepts, analyse relationnelle de concepts

Le besoin d'accès aux données par la société a conduit à la publication, par différents acteurs (gouvernement, universités, acteurs culturels), de vastes corpus de données exprimées dans les formalismes du web sémantique (principalement RDF).

Une part importante de la valeur ajoutée des données liées réside dans les liens identifiant la même entité dans différents jeux de données. Par exemple, cela permet d'identifier les mêmes ouvrages dans différentes sources de données bibliographiques. Les liens permettent d'exploiter conjointement les données de ces sources.

Par conséquent, la génération de tels liens est une tâche importante pour le web des données. Elle est en général pilotée par une spécification de liage. Différents types de spécifications sont disponibles. La plus répandue consiste à calculer une distance entre les identifiants de ressources et à estimer que plus ils sont proches, plus ils ont de chance d'identifier la même ressource.

Un autre type de spécification est ce que nous appelons *clé de liage*. Les clés de liage généralisent les clés de bases de données dans deux directions indépendances : elles fonctionnent avec des données représentés en RDF, et elles s'appliquent à deux jeux de données indépendants. Un exemple de clé de liage est :

$$\{\langle \mathsf{auteur}, \mathsf{creator}\rangle\}\ \ \{\langle \mathsf{titre}, \mathsf{title}\rangle\}\ linkkey\ \langle \mathsf{Livre}, \mathsf{Book}\rangle \tag{1}$$

qui signifie que si deux instances des classes Livre et Book respectivement, ont les mêmes valeurs pour les propriétés auteur et creator et au moins une valeur commune pour les propriétés titre et title, alors elles dénotent la même ressource.

Une première méthode a été conçue pour extraire les clés de liage entre deux classes (Atencia *et al.*, 2014). Elle commence par extraire des clés candidates puis les évalue à l'aide de mesures adaptées.

L'analyse formelle de concepts (FCA ou AFC) est une technique pour extraire des concepts entre deux ensembles ordonnés interdépendants (Ganter & Wille, 1999). L'analyse relationnelle de concepts (RCA) en est une extension permettant d'extraire des descriptions interdépendantes entre différents concepts (Rouane-Hacene *et al.*, 2013). L'AFC a déjà été utilisée pour extraire les clés dans le modèle relationnel.

L'étape d'extraction de clés candidates a été reformulée en un problème d'analyse de concepts formels pour le cas simple des bases de données (Atencia *et al.*, 2014). Nous avons étendu ce travail pour prendre en compte les attributs non fonctionnels et les dépendances entre clés de liage (lorsque les conditions d'une clé nécessitent de déterminer l'égalité de deux instances d'autres classes, ce qui utilise une autre clé, voir Table 1). Pour prendre en compte les références circulaires, il est devenu nécessaire d'adapter les techniques de RCA au cas des clés de liage.

| $K_{Person,Inhabitant}$ | $\forall\langle lastname, given\rangle$ | $\forall\langle lastname, name\rangle$ | $\exists\langle lastname, given\rangle$ | $\exists\langle lastname, name\rangle$ | $\forall\langle home, address\rangle_{C4}$ | $\exists\langle home, address\rangle_{C4}$ | $\forall\langle home, address\rangle_{C5}$ | $\exists\langle home, address\rangle_{C5}$ | $\forall\langle home, address\rangle_{C8}$ | $\exists\langle home, address\rangle_{C8}$ | $\forall\langle home, address\rangle_{C1}$ | $\exists\langle home, address\rangle_{C1}$ | $\forall\langle home, address\rangle_{C0}$ | $\exists\langle home, address\rangle_{C0}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\langle z_3, i_3\rangle$ | | × | | × | × | × | × | × | × | × | × | × | × | × |
| $\langle z_3, i_2\rangle$ | | × | | × | | | | | | | × | × | × | × |
| $\langle z_3, i_1\rangle$ | | | | | | | | | × | × | × | × | | |
| $\langle z_1, i_3\rangle$ | | | | | | | | | × | × | × | × | | |
| $\langle z_1, i_2\rangle$ | | | | | | | × | × | | | × | × | | |
| $\langle z_1, i_1\rangle$ | | × | | × | × | × | × | × | × | × | × | × | × | × |
| $\langle z_2, i_3\rangle$ | | × | | × | | | | | | | × | × | × | × |
| $\langle z_2, i_2\rangle$ | | × | | × | × | × | × | × | × | × | × | × | × | × |
| $\langle z_2, i_1\rangle$ | | | | | | | × | × | | | × | × | | |

| $K_{House,Place}$ | $\forall\langle city, city\rangle$ | $\exists\langle city, city\rangle$ | $\forall\langle owner, ownedBy\rangle_{C6}$ | $\exists\langle owner, ownedBy\rangle_{C6}$ | $\forall\langle owner, ownedBy\rangle_{C2}$ | $\exists\langle owner, ownedBy\rangle_{C2}$ | $\forall\langle owner, ownedBy\rangle_{C9}$ | $\exists\langle owner, ownedBy\rangle_{C9}$ | $\forall\langle owner, ownedBy\rangle_{C3}$ | $\exists\langle owner, ownedBy\rangle_{C3}$ | $\forall\langle owner, ownedBy\rangle_{C7}$ | $\exists\langle owner, ownedBy\rangle_{C7}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\langle h_1, a_2\rangle$ | | | | | × | × | | | × | × | | |
| $\langle h_1, a_1\rangle$ | × | × | × | × | × | × | × | × | × | × | × | × |
| $\langle h_1, a_3\rangle$ | | | | | | | | | × | × | × | × |
| $\langle h_3, a_2\rangle$ | × | × | | | | | × | × | × | × | | |
| $\langle h_3, a_1\rangle$ | | | | | | | | | × | × | × | × |
| $\langle h_3, a_3\rangle$ | × | × | × | × | × | × | × | × | × | × | × | × |
| $\langle h_2, a_2\rangle$ | × | × | × | × | × | × | × | × | × | × | × | × |
| $\langle h_2, a_1\rangle$ | | | | | × | × | | | × | × | | |
| $\langle h_2, a_3\rangle$ | × | × | | | | | × | × | × | × | | |

TABLE *1 – Contexte formel étendu après six itérations d'analyse relationelle de concepts conduisant au treillis de la Figure 1.*

Linkky [1] est un démonstrateur de ces techniques implémenté en Python 3 (Vizzini, 2017). Les bibliothèques RDFLib et Graphviz sont utilisées pour charger les graphes RDF et afficher les treillis de concepts respectivement (voir Figure 1). L'implémentation utilise l'algorithme de Norris (Norris, 1978) pour extraire les concepts. L'algorithme est étendu pour traiter des couples d'identifiants dans l'extant et des couples de propriétés quantifiées et qualifiées par la clé à utiliser pour la comparaison dans l'intant. Le processus d'analyse relationnelle de concepts est implémenté en appliquant itérativement deux opérateurs d'échelonnage.

Linkky prend en entrée deux jeux de données en RDF et retourne l'ensemble des clés candidates. Le système ne prend en compte aucun alignement a priori. Il utilise aussi les mesures développées dans (Atencia *et al.*, 2014) pour déterminer les familles de clés de liage candidates compatibles.

Le processus peut donc être résumé ainsi :

1. Charger les deux jeux de données RDF ;
2. Construire la famille de contextes relationnels ;
3. Appliquer FCA aux contextes formels ;

1. `http://moex.inria.fr/software/linkky/`

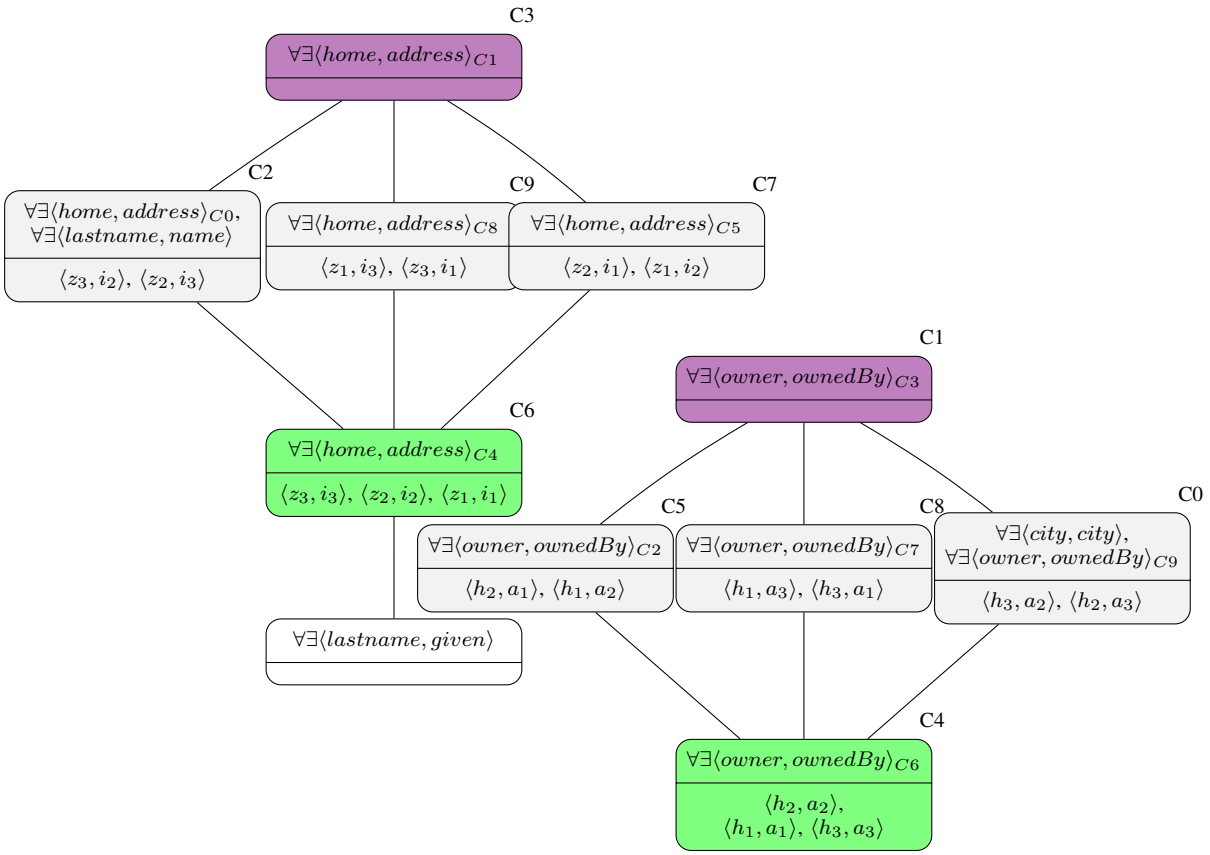



FIGURE *1 – Deux treillis de clés de liage candidates présentant des familles de clés interdépendantes (en vert et violet). Les deux clés vertes sont celles avec la meilleure évaluation (1. de couverture et de discriminabilité) et produisent le résultat espéré.*

4. Utiliser les opérateurs d'échelonnage pour introduire les nouveaux concepts créés dans les contextes formels ;
5. Si les contextes sont différents, aller en 3 ;
6. Extraire les familles de concepts compatibles (n'utilisant que des clés de la famille) ;
7. Évaluer la couverture et la discriminabilité des familles ainsi obtenues ;
8. Afficher contextes et treillis réduits.

Les différentes originalités de Linkky, outre d'être une implémentation de l'extraction de clés de liages en RDF, sont :

— il ne nécessite pas d'alignement entre les classes à considérer ;
— il peut extraire des clés entre différentes classes et une classe commune ;
— il extrait les familles de clés dépendantes ;
— il engendre directement l'affichage des treillis en LaTeX.

## Remerciements

Ce travail a été financé en parti par le projet ANR Elker (ANR-17-CE23-0007-01) pour les deux premiers auteurs et par une subvention du LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01) financé par le programme "Investissement d'avenir" pour Jérémy Vizzini.

## Références

ATENCIA M., DAVID J. & EUZENAT J. (2014). Data interlinking through robust linkkey extraction. In *Proc. 21st European Conference on Artificial Intelligence (ECAI)*, p. 15–20 : IOS Press.

ATENCIA M., DAVID J. & EUZENAT J. (2014). What can FCA do for database linkkey extraction ? In *Proc. 3rd ECAI workshop on What can FCA do for Artificial Intelligence ? (FCA4AI), Praha (CZ)*, p. 85–92.

GANTER B. & WILLE R. (1999). *Formal Concept Analysis*. Berlin : Springer.

NORRIS E. (1978). An algorithm for computing the maximal rectangles in a binary relation. *Revue Roumaine de Mathématiques Pures et Appliquées*, **23**(2), 243–250.

ROUANE-HACENE M., HUCHARD M., NAPOLI A. & VALTCHEV P. (2013). Relational Concept Analysis : mining concept lattices from multi-relational data. *Annals of Mathematics and Artificial Intelligence*, **67**(1), 81–108.

VIZZINI J. (2017). *Data interlinking with relational concept analysis*. Mémoire de master, Université Grenoble Alpes.