INFORSID 2017

Actes du congrès INFORSID

35e édition

Toulouse, 30 mai - 2 juin 2017

Editeurs : Olivier Teste, Nathalie Vallès-Parlangeau, André Péninou

L'association INFORSID

Siège Social : 44, Chemin de la Caille - 31750 Escalquens

Web : http://inforsid.irit.fr/

INFORSID est une association régie par la loi de 1901 qui rassemble les chercheurs en informatique des organisations et systèmes d'information et qui a pour objectif de promouvoir les recherches effectuées dans ces domaines en faisant intervenir le plus largement possible les utilisateurs et les industriels. INFORSID centre son activité sur un ensemble de colloques et de séminaires périodiques au cours desquels le point est fait sur l'état des recherches en matière de système d'information et une orientation est donnée pour leur prolongement.

Composition du bureau :

Présidente : Régine LALEAU, LACL, Université Paris-Est Créteil, IUT Sénart-Fontainebleau
Vice-président : Franck RAVAT, IRIT, Université Toulouse
Trésorier : Christian SALLABERRY, LIUPPA, Université de Pau et des Pays de l'Adour, IUT de Bayonne
Secrétaire : Agnès FRONT, LIG, Université Grenoble Alpes
Chargé de communication : Elöd EGYED-ZSIGMOND, LIRIS, Université de Lyon, INSA de Lyon

Présidents d'honneur :

Jean-Bernard CRAMPES (Toulouse)
Gilles ZURFLUH (Toulouse)
André FLORY (Lyon)
Claude CHRISMENT (Toulouse)
Michel SCHNEIDER (Clermont-Ferrrand)
Corine CAUVET (Aix-Marseille)
Chantal SOULE-DUPUY (Toulouse)
Dominique RIEU (Grenoble)

PRÉFACE

Les systèmes d'informations (SI) sont au cœur des grandes organisations. Ils remplissent différentes tâches de collecte, de stockage, de traitement et de mise à disposition de l'information au sein des entreprises et des administrations. Les SI couvrent ainsi un grand nombre de problématiques des sciences et technologies de l'information et de la communication.

Au cours des dernières décennies, la généralisation des technologies de l'information au travers de réseaux mondialisés, la diffusion massive de moyens de communications mobiles, et le développement d'objets autonomes connectés, permettent aux organisations mais aussi à chacun, de partager volontairement ou non, des données et des connaissances. L'humanité produit des quantités de données dans des proportions et avec un rythme sans commune mesure avec le passé. Ce nouvel environnement met en cause bon nombre d'approches classiques dans les systèmes d'information qui doivent faire face à des données très variables, structurées ou non, brutes ou plus élaborées, parfois imparfaites, non vérifiées, plus au moins persistantes, volatiles et dynamiques. Les masses de données disponibles aujourd'hui représentent des volumes disparates difficilement accessibles aux méthodes et outils traditionnels de collecte, de stockage et de gestion, d'exploitation, d'analyse et de restitution de l'information et de la connaissance.

Le congrès INFORSID (INFormatique des ORganisations et Systèmes d'Information et de Décision) tient sa 35e édition cette année. INFORSID rassemble chaque année, chercheurs et industriels sur l'ensemble des problématiques d'ingénierie et de gouvernance des systèmes d'information, de gestion des données, de leur manipulation et de leur exploitation.

Trois conférenciers invités, nous ont fait l'honneur d'accepter notre invitation. Il s'agit de Nicolas Monturet (Airbus), Camille Salinesi (CRI Paris I) et Zohra Bellahsene (LIRMM Montpellier).

Différents ateliers mettent le focus sur des problématiques variées : Open et/ou Linked Data dans les systèmes d'information ; surveillance et gestion de crise ; enseignement des SI ; valorisation et analyse des données de la recherche ; data visualisation dans les systèmes d'information ; systèmes d'information et de décision et démocratie ; combiner des données d'observation satellitaire avec d'autres sources pour l'aide à la décision et l'intelligence spatiale ; sécurité des systèmes d'information - technologies et personnes.

Cette année, le congrès INFORSID a reçu 45 soumissions d'articles couvrant un large spectre des problématiques liées aux systèmes d'information, des exigences à la mise en œuvre, des données aux processus métiers. Chacun des articles soumis a été évalué par trois membres du Comité de Programme. Une réunion plénière du Conseil du Comité de Programme a permis de sélectionner 20 articles acceptés pour présentation lors du congrès. Huit des articles acceptés ont fait l'objet d'un accompagnement par

un méta-relecteur pour aboutir à la version présente dans ces actes.

Avant de clôturer cette préface, je tiens à remercier les membres du bureau de l'association INFORSID, sous la présidence de Régine Laleau, pour m'avoir confié l'organisation scientifique du congrès, et pour leur assistance tout au long de cette année.

Je tiens également à remercier chaleureusement tous ceux qui ont contribué à l'organisation de ce congrès :

- en premier lieu les auteurs des 45 papiers pour leurs efforts dans la rédaction des articles et les conférenciers venus nous présenter les articles sélectionnés ;
- les membres du comité de programme qui par leurs relectures et l'attention portée à la rédaction de leurs évaluations, contribuent à dynamiser nos recherches ; les méta-relecteurs pour leur investissement dans l'aide à une meilleure diffusion des résultats de recherche ;
- les conférenciers invités pour avoir accepté de nous faire partager leur vision et leurs expériences ;
- les porteurs des ateliers pour leur investissement et leur dynamisme à organiser ces rencontres enrichissantes et originales ;
- les participants au congrès pour faire vivre et promouvoir notre communauté.

Enfin, je remercie le comité d'organisation, tout particulièrement Nathalie Vallès-Parlangeau et l'ensemble des jeunes doctorants qui ont préparé tout au long de cette année les journées INFORSID de Toulouse 2017. Je veux conclure cette préface par un dernier remerciement à André Péninou, qui a œuvré avec constance et efficacité à construire ces actes.

Olivier Teste
Président du comité de Programme INFORSID 2017

Ces journées ont été gérées en utilisant l'outil easychair.

COMITÉS

Le comité de la 35e édition d'INFORSID est composé par les responsables de l'organisation ainsi que les membres du comité de programme et les membres du conseil du comité de programme. Les président(e)s sont mentionné(e)s par une étoile (*).

Comité d'organisation INFORSID 2017

Amal Ait Brahim	Université Toulouse 1 Capitole, IRIT
Julien Aligon	Université Toulouse 1 Capitole, IRIT
Nadine Baptiste-Jessel	Université Toulouse 2 Jean-Jaurès, IRIT
Hamdi Ben Hamadou	Université Toulouse 3 Paul Sabatier, IRIT
Raphaëlle Bour	Université Toulouse 1 Capitole, IRIT
Max Chevalier	Université Toulouse 3 Paul Sabatier, IRIT
Maryse Colletis-Salles	Université Toulouse 1 Capitole, IRIT
Mohammed El Malki	Université Toulouse 1 Capitole, IRIT
Amir Laadhar	Université Toulouse 3 Paul Sabatier, IRIT
Christine Julien	Université Toulouse 3 Paul Sabatier, IRIT
Imen Megdiche	Université Champollion, IRIT
Manel Mezghanni	Université Toulouse 1 Capitole, IRIT
André Péninou	Université Toulouse 2 Jean-Jaurès, IRIT
Franck Ravat	Université Toulouse 1 Capitole, IRIT
Florence Sèdes	Université Toulouse 3 Paul Sabatier, IRIT
Jiefu Song	Université Toulouse 1 Capitole, IRIT
Chantal Soulé-Dupuy	Université Toulouse 1 Capitole, IRIT
Ronan Tournier	Université Toulouse 1 Capitole, IRIT
Nathalie Vallès-Parlangeau (*)	Université Toulouse 1 Capitole, IRIT

Conseil du Comité de Programme INFORSID 2017

Jean-Michel Bruel	Univ. Toulouse 2 Jean-Jaurès, IRIT
Corine Cauvet	Aix-Marseille Université, LSIS
Rebecca Deneckere	Université Paris 1 Panthéon Sorbonne, CRI
Dominique Rieu	Université Grenoble Alpes IUT2, UGA, LIG
Philippe Roose	Université de Pau et des Pays de l'Adour, LIUPPA
Chantal Soule-Dupuy	Université Toulouse 1 Capitole, IRIT
Gilles Zurfluh	Univ. Toulouse 1 Capitole, IRIT

Comité de programme INFORSID 2017

Ikram Amous Ben Amor	Enet'Com de Sfax, MIRACL
Eric Andonoff	Univ. Toulouse 1 Capitole, IRIT
Faten Atigui	CNAM Paris, Cedric
Henri Basson	Université Lille Nord, LISIC
Ladjel Bellatreche	ISAE-ENSMA, LIAS
Nicolas Belloir	Ecoles de Saint-Cyr - Coëtquidan, IRISA
Khalid Benali	Université de Lorraine, LORIA
Isabelle Borne	Université Bretagne Sud, IRISA
Emmanuel Bruno	Université de Toulon, LSIS
Max Chevalier	Université Toulouse 3 Paul Sabatier, IUT, IRIT
Adrian Chifu	Aix-Marseille Université, LSIS
Sophie Dupuy-Chessa	Université Grenoble Alpes, LIG
Faiza Ghozzi-Jedidi	Université de Sfax, Tunisie, MIREACL
Claude Godart	Université de Lorraine, LORIA
Akram Idani	INP Grenoble, LIG
Eric Kergosien	Université Lille 3, Geriico
Cyril Labbé	Université Grenoble Alpes, LIG
Sébastien Laborie	Université de Pau et des Pays de l'Adour, LIUPPA
Philippe Lamarre	INSA de Lyon, LIRIS
Anne Laurent	Université de Montpellier, Polytech, LIRMM
Myriam Lewkowicz	Univ. de Technologie de Troyes, ICD
Sabine Loudcher	Université Lyon 2, ERIC
Sofian Maabout	Université de Bordeaux, LABRI
André Miralles	IRSTEA Montpellier, IRSTEA
Isabelle Mirbel	Université Côte d'Azur, I3S
Elsa Negre	Univ. Paris Dauphine, LAMSADE
Noël Novelli	Aix-Marseille Université, LSIS
Verónika Peralta	Univ. de Tours, LI
Michaël Petit	Université de Namur, PReCISE
François Pinet	IRSTEA Clermont-Ferrand
Nicolas Prat	ESSEC Paris
Jolita Ralyte	Université de Genève, ISS
Philippe Ramadour	Aix-Marseille Université, LSIS
Mathieu Roche	Univ. Montpellier 2, CIRAD
Claudia Roncancio	INP Grenoble, LIG
Nurcan Selmin	Université Paris 1 Panthéon-Sorbonne, CRI
Dalila Tamzalit	Université de Nantes, IUT de Nantes, LS2N
Anne Tchounikine	INSA de Lyon, LIRIS
Olivier Teste (*)	Univ. Toulouse 2 Jean-Jaurès, IRIT
Virginie Thion	Université de Rennes 1, ENSSAT, IRISA
Ronan Tournier	Univ. Toulouse 1 Capitole, IRIT
Marlène Villanova	Université Grenoble Alpes, LIG

Relecteurs additionnels

Sofia Kleisarchaki, Pierre-Antoine Rappe, Selma Khouri, Therese Libourel, Paola Gomez, Maguelonne Teisseire, Mourad Bouneffa.

TABLE DES MATIÈRES

Conférences invitées

Sémantique des données et connaissances

Modèles : concepts et ingénierie

Filtrage d'informations

Processus : concepts et ingénierie

Patrons de conception

Analyse de l'information dans les réseaux sociaux

Gestion de données complexes

Ingénierie des méthodes

Conférences invitées

IoT in Airbus value chain

Nicolas Monturet[1]

1. Airbus
nicolas.monturet@airbus.com

Résumé.

The presentation will concern four main points:

- Introduction to Airbus
- Overview on IoT main business axis for Airbus
- Technology bricks and work streams
- Focus on Hangar of the future project

Nicolas Monturet is working for Airbus in IT Strategy, Enterprise Architecture and Innovation department. His job is about delivering technology watch, benchmarking, and architecture recommendations to contribute to Airbus Digital Transformation with a focus on IoT development. He is also supporting the development of airbus IoT center of competence as well as co-leading an Airbus IoT community.

Un jour, les Systèmes d'Information se concevront eux-mêmes

Camille Salinesi[1]

1. CRI, Paris1 Panthéon-Sorbonne
Camille.Salinesi@univ-paris1.fr

Résumé.

« Bonjour, je suis V2, votre nouveau logiciel d'entreprise. J'ai été conçu pour analyser vos besoins, et m'adapter seul de manière à vous assister au mieux dans vos activités et selon la situation. Comme V1, je peux reconfigurer dynamiquement mon paramétrage de manière complètement transparente, alignée avec la stratégie de votre entreprise, vos processus organisationnels et vos règles de gestion. Mieux, mon IA est capable d'enrichir mon code de manière entièrement autonome afin de permettre à votre entreprise d'offrir des services innovants. »

Quelle est la distance entre les systèmes d'information tels que nous sommes capable de les concevoir aujourd'hui, et V2 le logiciel intelligent qui devance vos besoins et s'y adapte entièrement seul ? Les progrès réalisés ces dernières années pour développer V2 sont déjà immenses : autrefois nous concevions les modèles de données, leur découverte automatique est aujourd'hui chose commune. Hier, la modélisation de processus était une tâche entièrement manuelle. Aujourd'hui, les techniques de fouille de processus automatisent la tâche. Il y a peu, une intelligence artificielle capable de générer son propre code a même été révélée au grand public !

Des éléments en provenance de différentes disciplines sont donc là, à notre disposition ou en passe de le devenir. Il existe même déjà des logiciels d'entreprise capables de s'adapter à leur environnement, et -dans une certaine mesure, d'évoluer. La communauté des Systèmes d'Information s'est elle-même emparée de certaines de ces avancées pour re-penser ses propres contributions. Un défi inédit se pose cependant: comment concevoir des Systèmes d'Information capables de s'auto-concevoir ?

Professeur des Universités, **Camille Salinesi** dirige le Centre de Recherche en Informatique de l´Université Paris 1 Panthéon – Sorbonne, une petite équipe d'accueil d'une trentaine de chercheurs spécialisée en Ingénierie des Systèmes d'Information et reconnue pour son expertise internationale en Ingénierie des

Exigences (IE). Camille a publié dans plus de 150 articles de revues et conférences internationales le fruits de ses recherches relatives à l'influence des méthodes, techniques, et outils de conception sur les qualités des systèmes à base de logiciels: satisfaction des utilisateurs, sécurité, réutilisabilité, durabilité, etc. Camille est notamment co-créateur de la méthode d'Ingénierie des Exigences à base de scénarios CREWS-L'Ecritoire, et de l'outil VARIAMOS d'Ingénierie de Lignes de Produits fondé sur la programmation par contraintes. Ses travaux les plus récents portent sur des approches innovantes de l'IE allant de la compréhension automatique des besoins des utilisateurs à la conception de logiciels auto-adaptatifs.

Investi dans la diffusion de la culture de l'IE, Camille Salinesi a contribué à la création du Master transdisciplinaire Informatique et Maitrise d'Ouvrage, a co-fondé l'association SPECIEF pour la promotion de l'IE en langue française, il co-anime le groupe de travail IE du GDR GPL, et est Vice Président du conseil exécutif de l'IREB, organisme international de certification en Ingénierie des Exigences. Son cours d'Ingénierie des Exigences est dispensé dans de nombreuses institutions et formations.

Rôle et techniques de l'alignement d'ontologies : un survol de l'état de l'art

Zohra Bellahsene[1]

1. LIRMM Montpellier
860 rue de St Priest, 34095 Montpellier cedex 5
Zohra.Bellahsene@lirmm.fr

Résumé.

Le besoin d'intégrer et d'analyser des grandes masses est présent dans de nombreux domaines d'applications. Le problème de l'alignement d'ontologies/schémas dont le résultat est un ensemble de correspondances entre différentes représentations du monde réel, est au centre du processus d'intégration des données. En effet, l'intégration de données est motivée par la forte hétérogénéité des données issues de sources multiples et l'absence de sémantique suffisante pour bien comprendre la signification des données. Citons le domaine biomédical où l'alignement d'ontologies joue un rôle clé dans le développement de la recherche biomédicale en facilitant le développement d'entrepôts de données articulés autour d'ontologies communes.

Cependant, les ontologies à aligner ont des structures différentes et n'utilisent pas le même vocabulaire (c'est-à-dire des termes différents pour décrire les mêmes concepts) parce qu'elles ont été conçues indépendamment par différents développeurs suivant différents principes et modèles. En outre, la diversité de leur hétérogénéité : syntaxique, terminologique (ou lexicale) et structurelle, ainsi que leur taille et leurs formats rendent la tâche d'alignement d'ontologie très difficile. L'alignement d'ontologie est un domaine de recherche actif en raison de son large éventail d'applications.

Lors de cette présentation, nous ferons un panorama des différentes approches et techniques sous-jacentes en les illustrant au travers d'outils d'alignement connus.

Zohra Bellahsene is a professor of Computer Science at University Montpellier and a senior researcher at LIRMM. She receive her PhD in CS from University of Paris 6 in 1982 and her HDR in CS from University Montpellier 2 in 2000. She has a long experience in database research and semantic Web, recently focusing on various

aspects of data integration, in particular, schema matching, view management and ontology matching. She has organized or chaired several international conferences and workshops, including being the PC co-chair of CoopIS2013, the PC chair of CAiSE'08, the co-chair of the XML Database Symposium (2006-2009), and the local chair of OTM06. She was the editor of the special issue of the DKE Journal on Data Integration over the Web in 2003. She was a coeditor of Schema and Matching and Mapping book, published in 2011 by Springer. She has been serving as PC member of major international conferences including VLDB, SIGMOD, ISWC, EDBT, ICDE, CAiSE, CIKM, ESWC, etc.

Sémantique des données et connaissances

LinkedMDR: un modèle sémantique de représentation de corpus de documents multimédia

Nathalie Charbel [1], Christian Sallaberry [2], Sébastien Laborie [1], Gilbert Tekli [4], Richard Chbeir [1]

1. UNIV PAU & PAYS ADOUR, LIUPPA, EA3000, 64600 ANGLET, FRANCE
{nathalie.charbel,sebastien.laborie}@univ-pau.fr,rchbeir@acm.org

2. UNIV PAU & PAYS ADOUR, LIUPPA, EA3000, 64000 PAU, FRANCE
christian.sallaberry@univ-pau.fr

3. UNIVERSITY OF BALAMAND (UOB), 100 TRIPOLI, LEBANON
gilbert.tekli@fty.balamand.edu.lb

RÉSUMÉ. *Dans le domaine du BTP, les projets de construction impliquent l'échange d'un volume important d'informations entre divers acteurs ayant des domaines d'expertises et des intérêts différents. La plupart des données échangées au sein de tels projets sont non ou semi-structurées, présentées dans des documents hétérogènes (souvent multimédia tels que des plans ou des rapports) et proviennent de sources variées. Bien évidemment, ces documents sont liés les uns aux autres par des liens explicites (p.ex., des références à tout ou partie de documents introduites par l'auteur) ou bien implicites (p.ex., selon les thèmes abordés dans les documents, tels que la plomberie, l'éléctricité ou l'isolation thermique du bâtiment). Identifier ce réseau de données liées entre documents tout au long de l'évolution d'un projet de construction, de l'indexation jusqu'à la recherche d'information, est aujourd'hui primordial pour faciliter la tâche d'un maître d'ouvrage ou d'un maître d'œuvre. Pour mener à bien cet objectif, dans cet article nous décrivons une nouvelle ontologie intitulée LinkedMDR (Linked Multimedia Document Representation). Cette ontologie est fondée sur l'intégration d'éléments issus de plusieurs standards de description de métadonnées multimédia dont Dublin Core (DC), Text Encoding Initiative (TEI) et Multimedia Content Description Interface (MPEG-7). Nous proposons de lier les standards de description les uns aux autres grâce à notre ontologie tout en fournissant de nouveaux concepts et relations non-pris en charge actuellement par ces standards. Cette représentation unifiée des documents nous permet donc de représenter sémantiquement un réseau de données liées sur un corpus documentaire. LinkedMDR est générique et offre une couche permettant de se spécialiser sur un domaine d'application métier (dans notre cas le BTP). Des expérimentations ont été menées afin de mesurer la qualité de notre proposition au regard d'autres solutions exploitant les standards de métadonnées multimédia actuels.*

ABSTRACT. *Projects, in the construction industry, involve the exchange of a large amount of information between several actors having different expertise and interests. Most of this information is unstructured, originated from different sources and dispersed across heterogeneous documents, thus producing implicit and explicit dependencies between them. This becomes very critical as it makes the annotation of the documents and the information retrieval more challenging at any stage of a building life cycle. In this work, we propose LinkedMDR: a novel ontology for Linked Multimedia Document Representation. Our ontology is based on the integration of the three standards addressing metadata and content representation: Dublin Core (DC), Text Encoding Initiative (TEI), and Moving Picture Experts Group (MPEG-7) together with the addition of new components offering more features especially in representing the collective knowledge of a document corpus. LinkedMDR is generic and offers, as well, a pluggable layer handling the particularities of a domain-specific knowledge. Experiments measure the efficiency and the effectiveness of our solution in comparison with the existing standards.*

MOTS-CLÉS : *Documents hétérogènes; Modèle de documents; Ontologies; Système de Recherche d'Information*

KEYWORDS: *Heterogeneous documents; Document Representation; Ontologies; Information Retrieval System*

1. Introduction

L'émergence des constructions durables, des architectures respectueuses de l'environnement économes en énergie ainsi que l'urbanisme moderne a conduit le domaine du BTP à suivre des approches communes de conception de projets immobiliers. Généralement, le processus de construction de tels projets est décomposé en trois phases essentielles : (i) la phase d'étude et de conception détaillée qui comprend la création, la préparation, l'analyse et la spécification des documents techniques et des plans d'exécution, (ii) la phase de construction et (iii) la phase opérationnelle qui couvre l'utilisation du bâtiment ainsi que la maintenance et le suivi du projet achevé (Klinger, Susong, 2006). Durant ce cycle de vie, de multiples acteurs (maître d'ouvrage, maître d'œuvre...) sont impliqués dans le processus de construction. Ils contribuent et échangent une grande variété de documents techniques et administratifs selon leurs expertises et leurs rôles au sein du projet. Par exemple, les contrats, les rapports techniques, les CCTP (Cahiers des Clauses Techniques Particulières), les CCAP (Cahiers des Clauses Administratives Particulières), les plans ainsi que les photos d'un projet sont généralement partagés durant les différentes phases de construction. Il apparaît très souvent que les documents échangés, provenant donc de sources différentes, n'ont pas de structure commune que ce soit entre les mêmes types de documents d'un projet (p.ex. des rapports techniques de type texte) ou que ce soit entre des documents similaires (p.ex., rapports thermiques) de projets de construction différents. Egalement, ils ont des versions, des formats d'encodage hétérogènes (p.ex., pdf, docx, xlsx, jpeg, etc.), des types de média différents (p.ex., images ou textes), évoquant des domaines métiers variés (p.ex., architecture, électricité, plomberie, mécanique, maçonnerie, etc.). De plus, ces documents peuvent comporter des références ainsi que des liens intra ou

inter-documentaires [1] de nature implicite ou explicite.
Dans la littérature, plusieurs travaux ont été engagés pour définir des métadonnées sur les documents et leur contenu. Ces modèles d'annotation peuvent être classés selon qu'ils traitent des contenus de type texte (e.g., TEI [2]), image (e.g., EXIF [3]) ou encore multimédia (e.g., (Arndt *et al*., 2007 ; Saathoff, Scherp, 2010 ; Garcia, Celma, 2005 ; Brut *et al*., 2009 ; Bloechle *et al*., 2006)). Néanmoins, aucun des travaux existants ne considère (i) un ensemble de documents multimédia hétérogènes, (ii) à la fois des annotations d'images et de textes qui portent sur leur contenu ainsi que leur structure, (iii) des spécificités liées à certains documents comme, par exemple, les légendes de plans, et (iv) différents liens inter et intra-documentaires. De plus, les standards actuels exploités dans le domaine du bâtiment, tel que l'IFC, se focalisent principalement sur des modèles de représentation d'objets 3D et ne prennent pas en considération des documents multimédia classiquement exploités durant un projet de construction.
Dans ce contexte, notre défi consiste à disposer d'une vue la plus claire et complète possible de ce réseau de données liées issu de documents multimédia relatifs à un projet immobilier. En effet, il est primordial de fournir aux différents acteurs un système d'information permettant de leur renvoyer des données souhaitées mais aussi et surtout de parcourir ce réseau de données en fonction de leur besoin et de leur expertise. Pour ce faire, nous devons au préalable définir un modèle qui permet de représenter ce réseau ainsi que la variété des données et des connexions qui le compose. Afin d'assurer l'interopérabilité des annotations, ce modèle supportera des concepts et des relations sémantiques exploitables lors de l'indexation, de la construction du réseau ou encore lors de la recherche d'information. En effet, la sémantique permettra aux différents acteurs de disposer d'informations pertinentes vis-à-vis de leurs expertises et de leurs préférences.
Dans cet article, nous proposons donc LinkedMDR : un modèle sémantique de représentation de réseau de données liées entre documents multimédia. LinkedMDR repose sur la combinaison d'éléments de standards de métadonnées existants tout en liant ces standards les uns aux autres, et en fournissant de nouveaux concepts et relations non-pris en charge actuellement par ces standards. Notre proposition est développée en collaboration avec la société Nobatek [4], une société française dans le secteur de la construction durable dont la mission consiste à assurer le transfert d'outils, de méthodes, de procédés et de produits innovants afin de contribuer à la performance énergétique et à la qualité environnementale. Dans ce cadre, des expérimentations ont été menées afin de mesurer la performance ainsi que l'efficacité de notre proposition au regard des autres solutions exploitant les standards de descriptions multimédia actuels. La présentation de notre contribution sera la suivante. La section 2 présentera plus en détail nos motivations au travers de situations réelles concrètes décrites par la société Nobatek.

1. Un lien inter-documentaire est un lien entre différents documents, tandis qu'un lien intra-documentaire est un lien entre deux éléments d'un même document.

2. Text Encoding Initiative, TEI P5 Guidelines for Electronic Text Encoding and Interchange, http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf.

3. Exchangeable image file format for digital still cameras, http://www.exif.org/Exif2-2.PDF.

4. http://www.nobatek.com

La section 3 fera état (i) des travaux existants dans le domaine de la représentation des métadonnées et des contenus, et (ii) des standards utilisés dans le domaine du bâtiment. Ces travaux ne satisfaisant pas des contextes métiers spécifiques, nous décrirons dans la section 4 notre ontologie nommée LinkedMDR et des résultats expérimentaux seront illustrés dans la section 5. La section 6 conclura cet article et développera de futures perspectives à envisager pour notre travail.

2. Motivations

Nous allons présenter un exemple de scénario de projets de construction qui nous a été fourni par la société Nobatek (§2.1). Ce scénario nous permet de dégager plusieurs défis qu'il conviendra de satisfaire par la suite (§2.2).

2.1. Le contexte

La société Nobatek est une société consultante qui assiste les maîtres d'ouvrages mais également les maîtres d'œuvres à développer leur projet de construction. Tout au long du cycle de vie d'un projet immobilier, cette société communique avec d'autres partenaires de construction, des bureaux d'études techniques, d'architecture, etc. Chaque acteur dispose donc de sa propre expertise que ce soit dans le domaine de la maçonnerie, de l'électricité, de la plomberie, etc. Dans ce contexte, de multiples documents (rapports techniques, plans...) sont élaborés faisant très souvent des références les uns aux autres. Par exemple, la figure 1 présente différents documents hétérogènes relatifs à un projet immobilier. Comme il est possible de le constater dans cette figure, il y a plusieurs rapports qui décrivent notamment les lots techniques du bâtiment (d_1 et d_5), ses propriétés thermiques (d_2) et acoustiques (d_3) ainsi qu'un extrait de plan d'étage du projet (d_4) et une photo (d_6). Actuellement, les ingénieurs de la société Nobatek doivent parcourir manuellement l'ensemble de ces documents. En effet, si ces derniers désirent vérifier la compatibilité des propriétés des façades extérieures avec les critères exigés par les normes environnementales, ils devront chercher d'eux-mêmes les informations dans ces documents rédigés par divers spécialistes (bureau d'étude thermique, bureau d'étude acoustique...). Cette recherche d'information est très fastidieuse et, de part le volume d'information important, peut conduire à ne pas consulter certains documents qui pourraient pourtant s'avérer utiles.

2.2. Les défis

– **Défi 1 : Représenter un réseau de données liées issu des documents** - Les ingénieurs doivent pouvoir rechercher et parcourir un réseau d'informations construit à partir d'un ensemble de documents. Il est évident de constater dans la figure 1 que les documents d_i ont de multiples relations les uns aux autres (p.ex., références, thématiques partagées, versions...). Ces relations peuvent être également de nature implicites ou explicites. Par exemple, certaines sections de texte entre d_1 et d_3 sont en relation implicite puisqu'elles décrivent la même thématique (p.ex., les façades extérieures). Le document d_5 quant à lui fait référence explicitement au plan technique

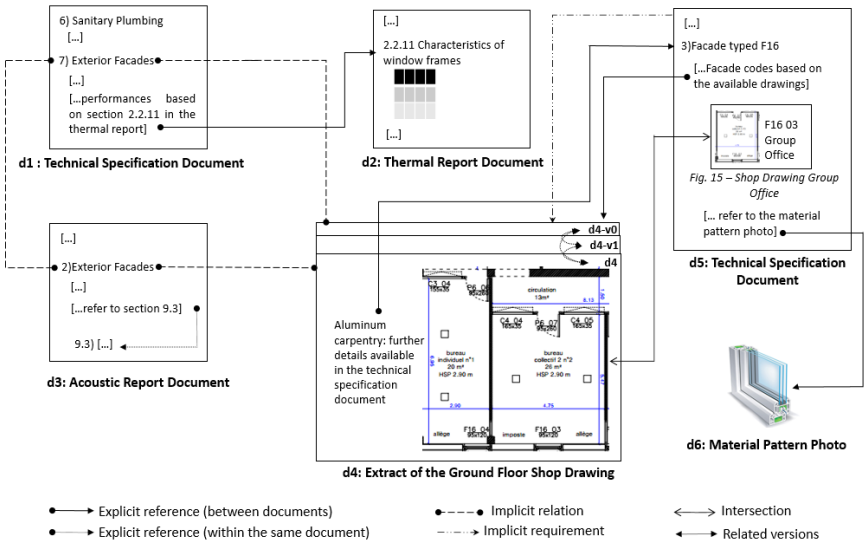


FIGURE *1. Exemple de documents multimédia hétérogènes relatifs à un projet de construction.*

d_4 (tout en sachant que de multiples versions de d_4 existent) et à la photo d_6. Par conséquent, il sera nécessaire d'indexer ces documents et d'en analyser les métadonnées afin d'établir ce réseau de relations.

– **Défi 2 : Exploiter la sémantique de l'information** - Les ingénieurs ont besoin de trouver avec différents niveaux de précision une information pertinente contenue au sein de documents évoquant de multiples thématiques. Par exemple, d_5 contient des données générales sur les façades extérieures, tandis que d_1 et d_2 contiennent des données plus détaillées sur ces façades extérieures avec d'autres éléments, tels que la plomberie et l'électricité. En effet, d_1 décrit les propriétés des façades comme pour la plomberie ou l'électricité, alors que d_2 se focalise particulièrement sur les propriétés thermiques des façades qu'elles soient intérieures ou extérieures. Outre la richesse des termes employés, il conviendra d'analyser les métadonnées d'ordre général ainsi que les structures des documents qui décrivent ces informations selon différents niveaux de granularité (p.ex., section, paragraphe ou phrase).

– **Défi 3 : Considérer la multimodalité des documents** - Les ingénieurs travaillent avec des documents hétérogènes ayant des types ainsi que des formats différents. Par exemple, d_1, d_2 et d_3 sont des documents Word, d_4 est un plan de construction, d_5 un document PDF et d_6 une photo au format JPG.

– **Défi 4 : Assurer l'extensibilité de l'information** - Les ingénieurs peuvent travailler sur d'autres types de construction avec d'autres types d'information et de média. Par exemple, dans un autre projet de construction que celui de la figure 1, ils peuvent

être amenés à analyser des documents sonores (de type audio) au sujet de bruits ambiants d'une pièce avant ou après la mise en place d'un revêtement spécifique sur les façades d'un bâtiment.

Dans ce qui suit, nous démontrons que les travaux de recherche actuels ne couvrent que partiellement ces défis au sein d'un même système d'information.

3. État de l'art

Nous allons présenter les standards et modèles existants de représentation de métadonnées, de structures et de contenus de documents multimédia (§3.1). De plus, nous décrirons un standard correspondant au domaine spécifique du bâtiment (§3.2). Nous conclurons cette partie par une comparaison ainsi qu'une discussion sur les limites de ces standards de description (§3.3).

3.1. Les standards et modèles existants de représentation de documents

Dublin-Core [5] - "Dublin Core Metadata Initiative" (DC) est un standard de métadonnées d'ordre général (p.ex., titre, date de création, format) décrivant une grande variété de documents. Il comprend 15 éléments ainsi que des composants, appelés "qualifiers", permettant le raffinement de ces éléments.

MPEG-7 [6] - "Multimedia Content Description Interface" est un standard décrivant différents types de contenu (p.ex., une image, une vidéo, un son). Il comprend trois composants principaux : les descripteurs (Ds) décrivant des éléments de base du contenu (p.ex., la couleur, la texture), les schémas de description (DSs) décrivant la structure et la sémantique des relations entre Ds et entre DSs, et le langage de définition de description qui est fondé sur les schémas XML (DDL).

Les modèles ontologiques - De nombreuses initiatives ont été menées pour spécifier des ontologies de description de documents multimédia. L'objectif principal de toutes ces approches consiste à combler le fossé entre les descripteurs de bas niveau, généralement extraits automatiquement par des indexeurs, et ceux de haut niveau exploités par les humains et décrivant la même information (Suarez-Figueroa *et al*., 2013). Comme indiqué dans les travaux de (Scherp *et al*., 2012), il est nécessaire de combiner plusieurs standards afin de disposer d'un système d'information multimédia le plus complet possible. Cette situation a donc ouvert la voie à la spécification d'ontologies dites multimédia. Par exemple, l'ontologie COMM (Core Ontology for MultiMedia) (Arndt *et al*., 2007) a été construite pour l'annotation de documents multimédia. Cette ontologie est fondée sur le standard MPEG-7, sur l'ontologie DOLCE ainsi que sur deux autres ontologies relatives aux design patterns. D'autres ontologies basées sur MPEG-7

5. Dublin Core Metadata Initiative, Metadata Basics, http://dublincore.org/documents/dcmi-terms/.

6. Multimedia content description interface, Technical report, Standard No. ISO/IEC n15938, 2001, http://mpeg.chiariglione.org/standards/mpeg-7/.

ont bien évidemment vu le jour, telles que MPEG-7 Rhizomik (Garcia, Celma, 2005) ou encore Multimedia Metadata Ontology (M3O) (Saathoff, Scherp, 2010). MPEG-7 Rhizomik fournit des correspondances directes entre le standard MPEG-7 et OWL, tandis que M3O propose un méta-modèle de représentation de documents multimédia qu'il est possible de spécialiser en fonction de son besoin. Le groupe de travail du W3C "Media Annotation Working Group" a également spécifié une ontologie intitulée "Media Resource Ontology" [7]. Cette dernière propose divers alignements avec les standards de métadonnées existants, tels que MPEG-7, Dublin Core et EXIF [8].

XCDF Format - XCDF est un format utilisé pour la représentation des résultats d'extraction et d'analyse des structures physiques de documents PDF (Bloechle *et al*., 2006). Ce format est basé sur le langage XML et sa DTD décrit un ensemble d'éléments permettant de représenter un document textuel : page, police, paragraphe, phrase, mot...

EXIF - Bien qu'il existe de multiples standards pour annoter des images, nous nous focaliserons sur le format EXIF (Exchangeable Image File Format) puisqu'il s'agit d'un format très complet permettant de décrire tout ou partie d'une image. En effet, ce langage comporte des éléments descripteurs de structure d'une image (hauteur, largeur, composition en terme de pixels), de version, de caractéristiques de l'image (couleurs, configuration, compression), d'informations sur son créateur (auteur, commentaires), sur le fichier (date de création, données GPS, droits...).

TEI - Il existe également de multiples standards de description de textes. Nous nous focaliserons sur le standard TEI (Text Encoding Initiative), basé sur XML. Le format TEI n'est pas seulement basé sur la structure du texte et ses annotations, il permet de faire référence à des concepts sémantiques qui facilitent la recherche d'information. Il est possible de classifier ces éléments selon les catégories suivantes : éléments sur la structure (p.ex., chapitres, sections, paragraphes, listes, tables), la mise en forme (polices de caractères), les annotations (titre, date, abréviations, signets, renvois), les figures et les graphiques.

3.2. Les standards de descriptions dans le domaine du bâtiment

Le standard IFC [9] (Industry Foundation Classes) est un des standards les plus utilisés pour l'échange de données BIM (Building Information Modeling) dans le domaine du bâtiment. Il contient toutes les informations utiles, comme les composants physiques d'un bâtiment, les espaces, les systèmes, les processus, les acteurs, ainsi que l'ensemble des relations entre ces éléments (Huovila, 2012). Les spécifications IFC peuvent être sérialisées en XML suivant un schéma XSD ou en EXPRESS, un autre langage de définition de données.

7. W3C, Ontology for media resource 1.0, http://www.w3.org/TR/mediaont-10/.

8. Exchangeable image file format for digital still cameras, http://www.exif.org/Exif2-2.PDF.

9. Industry Foundation Classes, IFC4 Add1 Release, http://www.buildingsmart-tech.org/specifications/ifc-releases/ifc4-add1-release.

3.3. Discussion

Nous avons étudié les standards et modèles existants de description de documents au regard du contexte ainsi que des défis que nous avons énumérés dans la section 2. Nos résultats sont synthétisés dans la Table 1.

TABLE *1. Synthèse des propriétés des standards et modèles existants au regard des défis dégagés.*

Défis	Standards et Modèles / Propriétés		Représentation de Métadonnées et de Contenus								
			Orientés Multimédia							Orientés Image	Orientés Texte
			Dublin Core	MPEG-7	COMM	M3O	MediaOnt	Mpeg-7 Rhizomik	Format Canonique XCDF	EXIF	TEI
Défi 1	Représentation d'un réseau sémantique de documents		x	x	x	x	x	x	x	x	x
	Description de relations	Intra-documentaire	Partielle	Partielle	Partielle	Partielle	Partielle	Partielle	x	x	Partielle
		Inter-documentaire	Partielle	Partielle	Partielle	Partielle	Partielle	Partielle	x	x	Partielle
Défi 2	Représentation de métadonnées descriptives		√	√	√	√	√	√	x	√	√
	Description de contenu		x	√	√	√	x	√	x	x	√
	Représentation de métadonnées de structure	Image	x	√	√	√	x	√	x	x	x
		Texte	x	Partielle	Partielle	Partielle	x	Partielle	Partielle	x	√
Défi 3	Multi-modalité		√	x	x	x	x	x	x	x	x
Défi 4	Extensibilité		√	√	√	√	√	√	√	x	√

Il est évident que les standards qui se focalisent exclusivement sur le texte ou bien les images sont limités puisqu'ils ne gèrent pas les différents types de documents multimédia (Défi 3). Néanmoins, ces standards, et notamment la TEI pour le texte, offrent des éléments de descriptions relativement complets et pertinents qu'il convient de réutiliser. Par exemple, la balise <*ref*> pour la TEI permet de faire des références mais celle-ci se limite au simple *confer*. Pour MPEG-7, il existe les éléments *Still Regions* et *Text Annotations* qui pourraient éventuellement servir à représenter globalement les différentes parties des plans techniques des bâtiments avec les légendes textuelles associées. Néanmoins, ces deux éléments ne permettent pas de représenter la structure complète des légendes avec différents niveaux de précision (Défis 2 et 4), ni même de faire des relations avec d'autres descripteurs provenant d'autres standards de représentation, comme la TEI ou DC (Défi 1).
Les langages et modèles de descriptions multimédia sont eux aussi limités. En effet, les ontologies et modèles multimédia agrègent simplement à l'aide de correspondances les descripteurs des standards existants mais n'apportent pas vraiment de plus-value (Défis 1 et 2).
En ce qui concerne les représentations dans le domaine du bâtiment, le standard IFC se focalise principalement sur la représentation 3D d'une construction. Il ne permet donc pas de relever tous les défis que nous avons exposé dans la section 2.2 (Défis 1 à 4).
D'autre part, il existe peu de travaux d'annotation de données multimédia respectant ces standards à des fins de recherche d'information (RI). Par exemple, le projet LINDO

(Large scale distributed INDexation of multimedia Objects) (Brut *et al*., 2009 ; 2011) a relevé le défi d'exploiter différents standards de métadonnées dans son système d'information distribué, tels que Dublin Core, EXIF et MPEG-7. Dans (Bates, 2011), l'auteur propose le langage CQL qui combine des recherches plein-texte et des recherches dans les métadonnées Dublin Core, à base de mots-clés. Enfin, dans (Feng *et al*., 2013), les auteurs présentent le potentiel du standard MPEG-7 à des fins d'annotation et de recherche d'information dans des corpus de documents multimédia.
De manière générale, à notre connaissance, il n'existe pas actuellement de représentation d'un réseau sémantique de données liées dans un corpus documentaire multimédia, ni de travaux de recherche d'information qui visent l'exploitation combinée de descripteurs de contenu thématique et de structure des documents, des relations intra et inter-documentaires, et des métadonnées plus générales. Notre proposition, illustrée dans la partie suivante, permettra de satisfaire ce besoin à travers l'ontologie LinkedMDR. Ce type d'approche est considéré comme le moyen le plus fiable et efficace pour supporter la recherche d'information sémantique à partir de données hétérogènes multimédias (Guo *et al*., 2017).

4. LinkedMDR : un modèle sémantique de représentation de corpus de documents multimédia

Nous proposons une ontologie, intituléé LinkedMDR, pour représenter un corpus de documents multimédia dans un seul modèle de données. LinkedMDR est basée sur (i) l'intégration des standards les plus pertinents en matière de représentation de méta-données et de contenus (notamment DC, TEI et MPEG-7) et (ii) l'ajout de nouveaux concepts et relations dépassant les limites de ces standards. Elle est constituée de trois couches principales : (i) la couche noyau servant de médiateur entre les différentes couches, (ii) la couche intégratrice de méta-données standards comprenant des éléments de standards existants et (iii) la couche spécifique au domaine, qui est adaptée à un domaine d'application particulier tel que le domaine de construction. L'idée de diviser l'ontologie en plusieurs couches et de centraliser les concepts et les relations les plus abstraits dans la couche noyau, assure sa généricité et son extensibilité (voir Figure 2).

4.1. La couche noyau

Cette couche comprend de nouveaux concepts et relations qui n'ont pas été adoptés par les standards existants, soient principalement : (i) les concepts qui modélisent la composition globale d'un document et les propriétés de méta-données qui lui sont associées (p.ex., *Document*, *Media*, *MediaComponent* et *DocumentProperty*), (ii) une entité *Object*, abstraction de *Document*, *Media* et *MediaComponent*, qui induit un riche ensemble de relations potentielles (relations *hasPart*, sémantique, temporelle et spatiale), (iii) des concepts qui généralisent ceux contenus dans la couche méta-donnée standard (*DescriptiveMetadata*, *AdministrativeMetadata* et *TextElement* généralisent des descripteurs de méta-donnée de DC et de structure de documents de TEI, respectivement) et (iv) de nouveaux concepts étendant le potentiel de description des standards MPEG-7 et TEI (*TextStillRegion* hérite de l'élément *TEI:Text* et l'élément

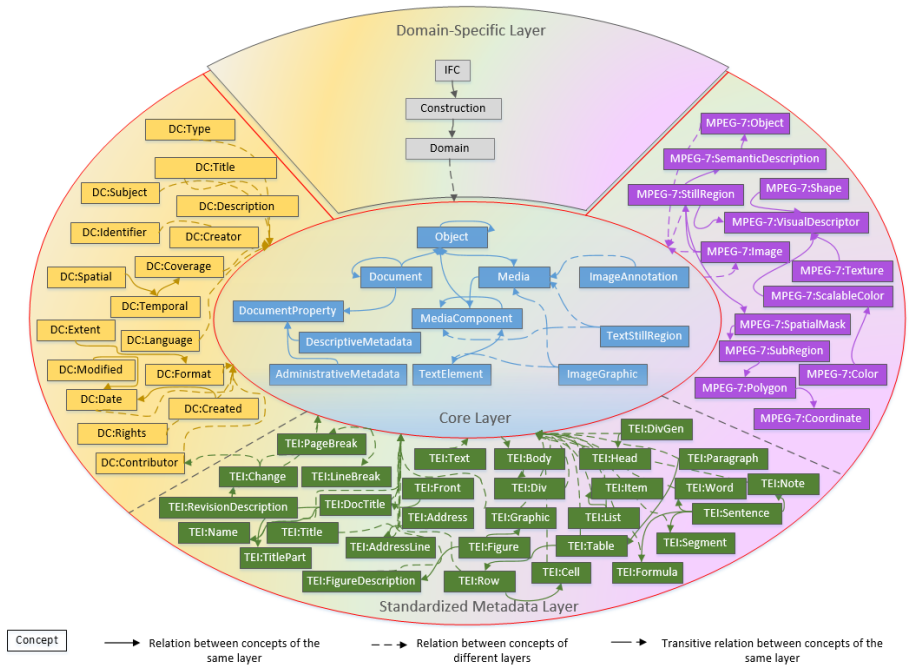


FIGURE 2. *LinkedMDR, le modèle sémantique de représentation de corpus de documents multimédia.*

MPEG-7:StillRegion pour représenter la structure d'une légende textuelle dans une image, *ImageGraphic* hérite de l'élément *TEI:Graphic* et l'élément *MPEG-7:Image* pour représenter les métadonnées d'une image insérée dans un texte et *ImageAnnotation* annote une image à l'aide d'une autre image ou d'un texte structuré).
Dans le reste du document, nous utilisons les triplets RDF [10] pour illustrer des instances de notre ontologie, car RDF est le modèle de données le plus largement utilisé pour représenter un graphe sémantique et extensible.
Ainsi, selon l'exemple de la figure 1, la section 3 (*div3*) du document d_5 contient une figure qui est un extrait de plan du document d_4. Cette relation se traduit par les triplets <*d5, hasPart, d5.div3.figure15*> et <*d4, contains, d5.div3.figure15*>. Cet exemple illustre une réponse apportée aux défis 1, 2 et 3 : il introduit une relation spatiale de contenance (*contains*) entre des documents de types différents. De plus, cet exemple illustre la sémantique attribuée à des niveaux de précision variés, ce qui permet d'inférer d'autres relations enrichissant le réseau de données (p.ex., si <*d5.div3.figure15, contains, dn.div1.figure1*>, on peut déduire que <*d4, contains, dn.div1.figure1*> par transitivité). Cela ne peut pas être réalisé avec les standards existants car ils ne permettent pas de représenter de telles relations à des niveaux de granularité différents.

10. Resource Description Framework, https://www.w3.org/RDF/.

4.2. La couche intégratrice de méta-données standards

Cette couche est constituée d'une sélection de méta-données définies par des standards : Dublin Core, TEI et MPEG-7. Par conséquent, elle se scinde en trois sous-couches, chacune dédiée à un standard. La première correspond à DC et comprend des méta-données d'ordre général relatives à un document. La deuxième présente les méta-données TEI décrivant la structure et le contenu d'un texte. Enfin, la troisième correspond aux méta-données décrivant une image avec ses différentes granularités, caractéristiques visuelles et descripteurs sémantiques suivant le standard MPEG-7. Il est à noter que, dans cet article, nous avons uniquement exploité les méta-données concernant les textes et les images. Cependant, à l'avenir, d'autres médias (comme audio et vidéo) pourront être considérés dans notre ontologie, en particulier dans cette couche.
Cette couche décrit également des relations entre ces différentes sous-couches. Par exemple, nous avons ajouté la relation *isRevisedBy* afin de relier le marqueur *TEI:Change* (décrivant l'ensemble de modifications apportées à un document) au marqueur *DC:Contributor* (celui qui a participé à ces modifications) correspondant. En outre, chaque sous-couche est également reliée à la couche noyau par l'intermédiaire de relations entre leurs concepts respectifs. Par exemple, nous pouvons citer <*TEI:Text, isA, Media*>, <*MPEG-7:StillRegion, isA, MediaComponent*>, <*DC:Title, isA, DescriptiveMetadata*>, <*TextElement, isOn, TEI:PageBreak*>.
Ainsi, selon l'exemple de la figure 1, le document d_4 a un ensemble de propriétés qui peuvent être décrites par la réutilisation des métadonnées de DC. À titre d'exemple, le titre de d_4 peut être traduit par les triplets <*d4, hasProperty, d4.title*> et <*d4.title, hasValue, "Shop Drawing"*>. De plus, d_4 contient des plans d'étage du bâtiment, chacun décrit sur une page du document. De ce fait, la réutilisation des méta-données du standard MPEG-7 peut servir à la description des différentes régions des plans techniques mais sans renseignement sur les pages correspondantes. De même, la réutilisation des méta-données de la TEI sert à la description de la répartition des plans sur les pages du document mais sans information relative au contenu de ces mêmes plans. Ainsi, dans le cadre du défi 1, la liaison entre les méta-données de ces différents standards n'est possible qu'à travers les concepts de la couche noyau comme le montrent les triplets suivants : <*d4, hasPart, d4.imagegraphic*>, <*d4.imagegraphic, isOn, d4.page1*>, <*d4.imagegraphic, hasPart, d4.stillregion*>.

4.3. La couche spécifique au domaine

Bien que notre proposition soit faite dans le contexte du domaine de la construction, nous visons à fournir une ontologie générique qui pourrait être utilisée dans n'importe quel domaine spécifique. Les couches mentionnées précédemment sont génériques et indépendantes du contexte dans lequel les documents multimédia sont utilisés. Toutefois, il est important de tenir compte d'éventuelles particularités relatives à un domaine spécifique. Par conséquent, nous présentons cette couche comme une illustration de la façon dont nous pouvons mettre en œuvre cette ontologie générique tout en l'adaptant à une utilisation ciblée, comme le domaine de la construction, par exemple.

Pour ce faire, nous présentons un nouveau concept intitulé *Domain* et nous le lions au concept *Object* de la couche noyau. De cette façon, des concepts spécifiques à un domaine donné peuvent être ajoutés sous *Domain* et par la suite seront en relation avec les sous-concepts de *Object* (c'est-à-dire, *Document*, *Media* et *MediaComponent*).
Dans cet article, nous présentons un exemple montrant comment nous pouvons rendre cette couche adaptable au domaine de la construction. Nous ajoutons le concept *Construction* comme sous-concept de *Domain*. Nous relions également ce dernier au concept *IFC* qui comprend les concepts de ifcOWL [11], la conversion du standard IFC en ontologie.
À titre d'exemple, la section 7 (*div7*) de d_1 (cf. Figure 1) décrit les façades extérieures. Il est maintenant possible de lier la section 7 avec l'objet IFC correspondant (p.ex., ifcwindow4): <*d1, isA, Document*>, <*d1, hasPart, d1.div7*>, <*d1.div7, isA, TEI : Div*>, <*ifcwindow4, isA, IFC: BuildingElement*> et <*ifcwindow4, isRelated, d1.div7*>. Cela répond particulièrement aux défis 1 et 4.
Pour plus de détails sur l'ontologie LinkedMDR, les différents concepts et relations appartenant à chacune des couches décrites dans cette partie sont disponibles en ligne avec la documentation correspondante : http://spider.sigappfr.org/linkedmdr/.

5. Expérimentation

Nous avons expérimenté l'annotation de documents hétérogènes, liés aux projets de construction dans l'entreprise Nobatek, afin d'évaluer deux critères : (i) la performance de notre modèle de données en terme de concision des annotations et (ii) son efficacité (qualité de l'annotation – rappel, précision, F_1-Mesure) par rapport aux standards existants en matière de représentation de méta-données et de contenu, plus particulièrement DC, TEI et MPEG-7. Nous décrivons tout d'abord le jeu de données de test puis nous commentons les résultats d'expérimentation.

5.1. Données de test

Nous avons sélectionné 6 documents relatifs à des projets de construction présentés dans le scénario de motivation initial (cf. Figure 1). Bien que ce nombre paraisse réduit, ces documents sont choisis à la main et peuvent représenter un scénario complet qui met en valeur tous les défis déjà dégagés (cf. Section 2.2). Nous avons ensuite effectué les cinq expérimentations suivantes :

– **Test#1: Annotation selon DC**

Nous avons utilisé Dublin Core Advanced Generator [12] afin de générer, pour chaque document, une représentation XML correspondant au standard DC. Ce test a permis de générer un fichier d'annotation XML pour chaque document.

11. http://ifcowl.openbimstandards.org/IFC4_ADD1.owl
12. Disponible en ligne sur http://www.dublincoregenerator.com/generator.html

TABLE *2. Évaluation de la concision des annotations dans les différents jeux de tests.*

Groupes de Tests	Nb. de Documents Annotés	Nb. Cumulé d'Annotations	Nb. de Fichiers XML Générés	Nb. de Redondances
Test#1	6	79	6	0
Test#2	4	646	4	0
Test#3	2	198	2	0
Test#4	6	923	12	91
Test#5	6	656	1	0

– **Test#2: Annotation selon TEI**
Nous avons utilisé l'outil OxGarage [13] afin de générer, pour chaque document, une représentation TEI P5 XML correspondant au standard TEI. Puis, nous avons ajouté à la main des références internes et externes (éléments *ptr* et *ref*) pour compléter les annotations. Ce test a permis de générer un fichier d'annotation XML pour chaque document textuel (c'est-à-dire d_1, d_2, d_3 et d_5).

– **Test#3: Annotation selon MPEG-7**
Nous avons utilisé l'outil Caliph V0.9.27 [14] afin de générer, pour chaque document, une représentation XML correspondant au standard MPEG-7. Nous avons ensuite apporté quelques modifications manuelles : nous avons ajouté des éléments qui n'étaient pas décrits par l'outil Caliph V0.9.27 (par exemple, *FreeTextAnnotation* associé à des éléments *StillRegion*). Ce test a permis de générer un premier fichier d'annotation XML pour l'image (d_6) et un second pour l'extrait de plan (d_4).

– **Test#4: Annotation selon DC, TEI et MPEG-7 (combinés)**
Nous avons utilisé les résultats des annotations issues de Test#1, Test#2 et Test#3.

– **Test#5: Annotation selon LinkedMDR**
Nous avons créé des instances de l'ontologie LinkedMDR [15] via Protégée pour construire un fichier RDF représentant tous les documents selon ce modèle.

5.2. Résultats d'expérimentation

5.2.1. Évaluation de la performance

Nous avons évalué la performance sur le plan de la concision des annotations générées par les tests, en considérant l'ensemble des six documents.
Ainsi, nous comparons les annotations fournies par les cinq tests en termes de (i) nombre cumulé d'annotations [16] ; (ii) nombre de documents source annotés ; (iii) nombre de fichiers XML générés ; (iv) nombre de redondances (chevauchement de méta-données). L'objectif est ici de mettre en valeur le scénario qui génère le nombre minimum d'an-

13. Disponible en ligne sur http://www.tei-c.org/oxgarage/. Cet outil ne traite pas les documents PDF, nous avons donc utilisé le service Web PDF to DOCX disponible à l'adresse http://pdf2docx.com/ pour convertir des documents PDF en DOCX.

14. Disponible sur http://www.semanticmetadata.net/

15. Disponible sur http://spider.sigappfr.org/download/1175/

16. Le nombre de balises XML dans les fichiers d'annotation ou le nombre de triplets dans l'ontologie.

notations (sans perte d'information), de fichiers XML et de redondances. Ces résultats sont présentés dans la table 2. Nous constatons que seuls Test#1 (DC), Test#4 (DC, TEI, MPEG-7) et Test#5 (LinkedMDR) ont été en mesure de générer des annotations pour chacun des six documents. DC montre un nombre faible d'éléments d'annotation mais ne couvre que les méta-données génériques sans considérer la structure et le contenu des documents. LinkedMDR, quant à elle, couvre l'ensemble du potentiel d'annotation attendu et, pour autant, génère un nombre réduit d'annotations.
En ce qui concerne les annotations résultant de Test#4 et de Test#5, la table 2 montre de bons résultats pour LinkedMDR puisque ce scénario de test a permis de représenter, dans un même fichier d'annotation, les six documents avec un nombre relativement réduit d'éléments d'annotation et sans redondance de méta-données. Test#4, quant à lui, génère un nombre important d'annotations, parce que TEI et MPEG-7 sont très verbeux, sans toutefois couvrir l'ensemble du potentiel d'annotation attendu.
Les résultats d'annotation seront ensuite exploités à des fins de recherche d'information (RI). Des éléments de méta-données, de structure et de contenus représentés de façon concise, sans redondance, dans un seul document de description, seront d'un intérêt majeur dans des scénarios de RI. Par exemple, pour la requête *« Quels contenus, issus de cahiers de clauses, traitent de façades et font référence à des plans d'étage ? »*, nous pouvons interroger les triplets RDF faisant référence à notre ontologie LinkedMDR. Ces triplets nous permettent de retrouver la section 3 (div 3) du document d_5 puisque d_5 est un CCTP (cahiers de clauses), traite de façades extérieures (façades) et inclut d_4 (plan). Ceci n'est pas possible avec les annotations du Test#4, l'information y est incomplète et répartie de façon indépendante dans plusieurs fichiers d'annotation, selon différents standards.

5.2.2. Évaluation de l'efficacité

Nous avons également évalué l'efficacité de ces modèles en calculant les scores de Précision, Rappel et F_1-Mesure relatifs aux annotations générées pour chacune des séries de tests. Puisque les éléments d'annotation varient entre les standards et LinkedMDR, nous avons fixé un ensemble de critères pertinents sur lesquels nous avons fondé nos calculs indépendamment du nombre d'annotation ou de leur nature (balises XML ou triplets RDF). Ces critères se scindent en plusieurs catégories : liens sémantiques inter ou intra-documentaires, liens topologiques inter-documentaires, méta-données générales ou spécifiques aux textes/images, etc. Nous mesurons des scores de précision [17] et de rappel [18] et de f_1-mesure [19] relatifs à chacun des groupes de test selon ces critères. Dans cette évaluation, nous avons également utilisé deux documents supplémentaires : un fichier audio et une vidéo, relatifs à des constructions, qui sont sémantiquement liées à des documents de notre jeu de tests. Les résultats des tests sont présentés dans la figure 3. De façon générale, l'annotation selon le standard TEI (Test#2) offre des résultats plus efficaces que ceux correspondants à DC (Test#1) ou

17. Nombre de critères pertinents couverts pour le test / Nombre total de critères annotés par le test.
18. Nombre de critères pertinents couverts pour le test / Nombre de critères pertinents attendus pour le test.
19. (2×P×R) /(P+R)

MPEG-7 (Test#3). Même lorsque les trois sont combinés, le score de F_1-Mesure, qui combine précision et rappel, ne change que légèrement de 0.59 (Test#2) à 0.61 (Test#4). Ceci en raison de la présence d'un grand nombre de méta-données textuelles annotées sur la base du standard TEI et d'un certain nombre de redondances avec les autres jeux d'annotations. En outre, nous ajoutons que la présence de relations structurelles entre composants textuels et documents a fortement contribué à la mise en valeur du Test#2 (TEI). LinkedMDR s'avère être le plus efficace avec le meilleur score de F_1-Mesure qui a atteint environ 0.94 (Test#5). Seuls les types de documents audio et vidéo ne sont pas représentés par LinkedMDR. Bien que ces types de document ne soient pas dans les objectifs d'annotation visés pour l'instant, ils peuvent l'être dans le futur (défi 4). Notre ontologie, qui est basée sur MPEG-7, est facilement extensible pour couvrir ultérieurement ces types de documents.

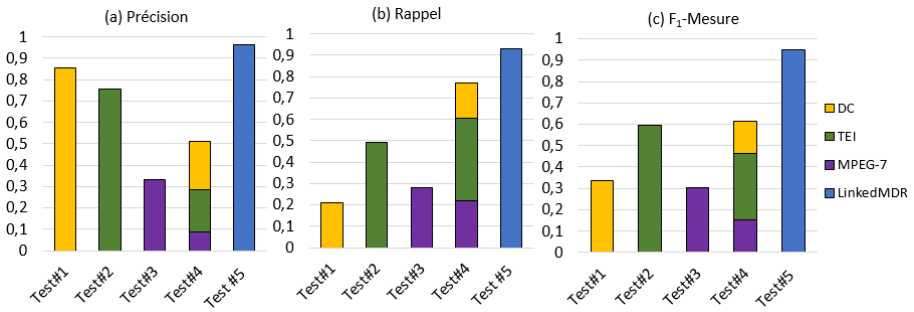


FIGURE *3. Évaluation de l'efficacité des modèles d'annotations.*

6. Conclusion

Cet article présente LinkedMDR, une nouvelle ontologie décrivant un corpus documentaire multimédia. LinkedMDR est adapté à tout corpus de documents hétérogènes spécifiques à un ou plusieurs domaines. Cette ontologie est basée sur l'intégration d'éléments des standards DC, TEI et MPEG-7 complétée par l'introduction de nouveaux concepts et relations dépassant leurs limites de représentation. LinkedMDR se scinde en plusieurs couches qui montrent explicitement, d'une part, sa généricité et, d'autre part, son potentiel de spécialisation, chacune mettant en exergue des possibilités d'extensions futures. Les expérimentations montrent de bons résultats de mesure de performance et d'efficacité d'annotations de méta-données et de contenus basées sur l'usage de l'ontologie LinkedMDR en comparaison aux annotations obtenues avec l'usage de standards existants.
Actuellement, nous développons une chaine de traitement automatique pour annoter un corpus de documents hétérogènes selon l'ontologie LinkedMDR et montrer que notre proposition peut être mise en place dans des scénarios réels. Pour ce faire, nous sommes en train d'exploiter les techniques avancées de collecte et d'extraction de métadonnées (Greenberg, 2004) ainsi que les techniques de traitement automatique du langage naturel comme dans (Maynard *et al.*, 2016). À court terme, nous envisageons également d'étendre nos expérimentations par des jeux de documents plus importants

et visons de nouveaux résultats confirmant une nouvelle fois la validation de notre modèle.

Bibliographie

Arndt R., Troncy R., Staab S., Hardman L., Vacura M. (2007). *COMM: designing a well-founded multimedia ontology for the web*. Springer.

Bates M. J. (2011). *Understanding information retrieval systems: management, types, and standards*. Auerbach Publications.

Bloechle J.-L., Rigamonti M., Hadjar K., Lalanne D., Ingold R. (2006). XCDF: a canonical and structured document format. In *International workshop on document analysis systems*, p. 141-152. Springer.

Brut M., Codreanu D., Manzat A.-M., Sèdes F. (2011). Distributed multimedia indexing and optimal resources utilization: An implementation based on metadata, context and usage. *JMPT*, vol. 2, n° 4, p. 197–225.

Brut M., Laborie S., Manzat A.-M., Sedes F. (2009). Integrating heterogeneous metadata into a distributed multimedia information system. *COGnitive systems with Interactive Sensors*.

Feng D., Siu W.-C., Zhang H. J. (2013). *Multimedia information retrieval and management: Technological fundamentals and applications*. Springer Science & Business Media.

Garcia R., Celma O. (2005). Semantic integration and retrieval of multimedia metadata. In *5th international workshop on knowledge markup and semantic annotation*, p. 69-80.

Greenberg J. (2004). Metadata extraction and harvesting: A comparison of two automatic metadata generation applications. *Journal of Internet Cataloging*, vol. 6, n° 4, p. 59–82.

Guo K., Liang Z., Tang Y., Chi T. (2017). Sor: An optimized semantic ontology retrieval algorithm for heterogeneous multimedia big data. *Journal of Computational Science*.

Huovila P. (2012). Linking IFCs and BIM to sustainability assessment of buildings. In *Proceedings of the cib w78 2012: 29th international conference.*

Klinger M., Susong M. (2006). The construction project: phases, people, terms, paperwork, processes. In, chap. Phases of the Contruction Project. American Bar Association.

Maynard D., Bontcheva K., Augenstein I. (2016). Natural language processing for the semantic web. *Synthesis Lectures on the Semantic Web: Theory and Technology*, vol. 6, n° 2, p. 1–194.

Saathoff C., Scherp A. (2010). Unlocking the semantics of multimedia presentations in the web with the multimedia metadata ontology. In *Proceedings of the 19th international conference on world wide web*, p. 831-840. ACM.

Scherp A., Eissing D., Saathoff C. (2012). A method for integrating multimedia metadata standards and metadata formats with the multimedia metadata ontology. *International Journal of Semantic Computing*, vol. 6, n° 01, p. 25-49.

Suarez-Figueroa M. C., Atemezing G. A., Corcho O. (2013). The landscape of multimedia ontologies in the last decade. *Multimedia tools and applications*, vol. 62, n° 2, p. 377-399.

Projet ModRef : Migration de Données vers des Triplestores CIDOC-CRM

Pascaline Tchienehom [1]

Université de Paris 10 - Labex "Les passés dans le présent",
200 Avenue de la République, 92000 Nanterre, France
pkenfack@u-paris10.fr

ABSTRACT. *ModRef is a project from the laboratory Labex "Les passés dans le présent", which coordinates various projects on digital humanities. ModRef focuses more precisely on the semantic web and linked open data. The goal is to move heterogeneous data into triplestores also called data warehouses or collections of RDF files in order to improve the sharing, exchange and discovery of new knowledge. For this purpose, the CIDOC-CRM norm has been chosen since it is, at present, the reference for the semantic description of museographic or cultural heritage data. In order to realise the proof of concept of ModRef, a general architecture has been defined, a semantic modelling and data mapping of selected sub-projects of ModRef have been proposed, triplestores have also been created. A web application has been implemented and deployed. This web application describes the ModRef project, as well as it enables visualising, querying and exploring created triplestores.*

RÉSUMÉ. *ModRef est un projet du laboratoire Labex "Les passés dans le présent" qui accompagne divers projets sur des problématiques relatives aux humanités numériques. Le projet ModRef s'intéresse spécifiquement au web sémantique et aux données ouvertes et liées. Le but de ce projet est de réaliser une migration de données hétérogènes vers des triplestores encore appelés entrepôts ou collections de fichiers RDF afin d'améliorer le partage, l'échange et la découverte de nouvelles connaissances. Pour ce faire, la norme CIDOC-CRM a été choisie car elle est actuellement la norme de référence pour la description sémantique de l'information muséographique ou d'héritage culturel. Afin de réaliser la preuve conceptuelle de ModRef, une architecture générale a été définie, une modélisation sémantique et un alignement des données des trois sous projets pilotes de ModRef ont été proposés, une migration des données vers des triplestores a également été effectuée. Une application web a été développée et déployée. Cette application web décrit le projet ModRef et permet également de consulter et d'interroger les triplestores créés.*

KEYWORDS: *Digital Humanities, Semantic Web, Triplestores, CIDOC-CRM, Linked Open Data.*

MOTS-CLÉS : *Humanités Numériques, Web Sémantique, Triplestores, CIDOC-CRM, Données ouvertes et liées.*

1. Introduction

Le Labex "Les passés dans le présent" accompagne de nombreux projets en Sciences Humaines et Sociales (SHS) sur des problématiques relatives aux humanités numériques (Oldman *et al.*, 2014) : de la dématérialisation des données à la description structurée voire sémantique de ces dernières. Le projet ModRef (Modélisation, Référentiels et Culture Numérique) du Labex fédère un ensemble de sous projets pour réaliser une migration de leurs données vers des triplestores encore appelés entrepôts ou collections de fichiers RDF (Resource Description Framework) afin d'améliorer le partage, l'échange et la découverte de nouvelles connaissances. Pour ce faire, la norme CIDOC-CRM (International Committee for DOCumentation - Conceptual Reference Model) (Boeuf *et al.*, 2015) a été choisie car elle est aujourd'hui la norme de référence pour la description sémantique des informations muséographiques ou d'héritage culturel (Hooland, Verborgh, 2014). Il s'agit généralement de passer de données non structurées ou semi structurées vers des données structurées puis vers des données sémantiques. Le web sémantique propose une solution pour réaliser ces migrations.

Le web sémantique (Shadbolt *et al.*, 2006) (Berners-Lee *et al.*, 2001) n'est pas qu'un concept mais également une architecture validée et de plus en plus éprouvée sous la forme d'un ensemble de couches indépendantes mais qui s'interfacent les unes avec les autres pour réaliser différentes tâches. Cette architecture décrit les données de leur représentation à leur exploitation via des applications ou agents web sémantique. Ainsi, de nombreuses normes de représentation de données pour le web sémantique existent. Le CIDOC-CRM est un exemple de norme sémantique et est plus spécifiquement un modèle conceptuel de référence ou une ontologie. Le but de la sémantique et donc des nombreux langages de métadonnées ou normes qui la composent est de fournir un cadre homogène de description et d'interrogation de sources de données hétérogènes afin de réduire le silence informationnel et d'améliorer la découverte de connaissances. Ainsi, le projet ModRef a pour but de réaliser une migration de données vers des triplestores CIDOC-CRM en s'appuyant sur des sources de données hétérogènes tant sur le contenu que sur la structure logique initiale (tableurs, bases de données relationnelles, fichiers XML).

Dans cet article, nous présentons le projet ModRef au travers : d'une description générale de la norme CIDOC-CRM ; de l'architecture du projet ModRef ; de la modélisation sémantique CIDOC-CRM et de l'alignement des données des trois sous projets pilotes de ModRef ; de la migration des données vers des triplestores CIDOC-CRM ; de la visualisation et de l'exploitation des triplestores avec l'application web http://triplestore.modyco.fr qui a été développée et déployée ; d'une procédure d'évaluation et des résultats obtenus.

2. Présentation générale de la norme CIDOC-CRM

Il existe plusieurs modèles de représentation de données basés sur de la sémantique (Scarinci, Myers, 2014) qui utilisent des langages de métadonnées qui décrivent des concepts et/ou des liens entre concepts ou instances de concepts (Dublin Core, RDF,

RDFS, OWL, FOAF, Wordnet, CIDOC-CRM). Le *CIDOC-CRM* (Boeuf *et al.*, 2015) (cf. http://www.cidoc-crm.org/) est un modèle conceptuel de référence pour l'information muséographique ou d'héritage culturel. La version de la norme CIDOC-CRM qui a été utilisée est la version 6.2 de mai 2015. Elle comporte 94 classes et 168 propriétés. Il faut noter que les travaux sur le CIDOC-CRM ont débuté en 1996 et c'est en 2006 que le CIDOC-CRM est devenu une norme ISO 21127. Cette norme permet de décrire les caractéristiques globales des objets (identifiant, type, titre, matériau, dimension, note) mais également leur historique au travers des évènements ou activités (transfert de garde -localisations anciennes, localisation actuelle-, origine, découverte, conservation, affectation de valeur, mesure) ainsi que les relations qui existent entre objets ou parties d'objets (bibliographie, composition, similarité, autre représentation -photo, dessin, tableau-, inscription). Une implémentation OWL (Ontology Web Language) du CIDOC-CRM de l'Université d'Erlangen-Nuremberg est disponible à l'adresse suivante : http://www.erlangen-crm.org/. Notons que l'espace de nom de cette implémentation du CIDOC-CRM est généralement préfixé par "*ecrm*".

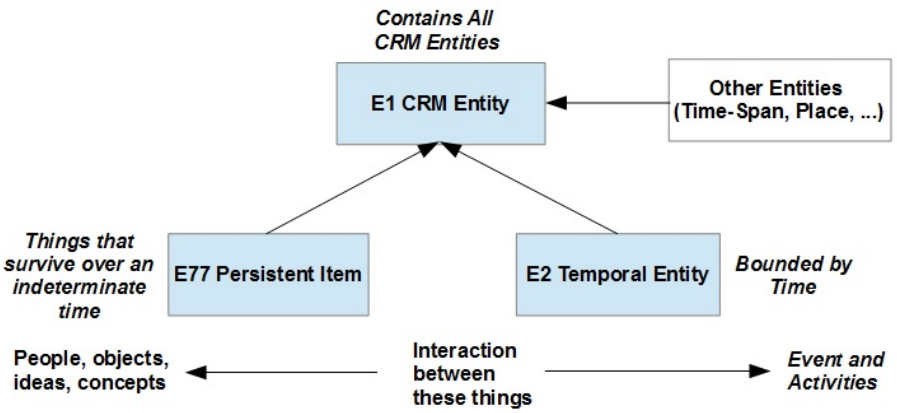


Figure 1. Structure générale des entités du CIDOC-CRM.

La structure générale du CIDOC-CRM est celle de la figure 1. La classe mère de toutes les entités du CIDOC-CRM est la classe *E1 CRM Entity* et elle se subdivise en sous-classes directes dont les deux principales sont : (1) *E77 Persistent Item* qui est la classe la plus générique des entités dites persistantes. Une entité persistante est une entité qui est capable de survivre pendant une période indéterminée, comme par exemple : les personnes, les objets, les idées, les concepts. Ce sont généralement des entités pouvant avoir un début ou une fin d'existence (destruction, par exemple) ; (2) *E2 Temporal Entity* qui est la classe la plus générique des entités dites temporelles. Une entité temporelle est une entité qui est limitée dans le temps, comme : un évènement, un début d'existence, une fin d'existence, une activité, une création, une production, une modification, un transfert de garde, une conservation, une mesure.

Les autres sous-classes directes de la classe racine *E1 CRM Entity* sont les classes *E52 Time-Span, E53 Place, E54 Dimension, E92 Spacetime Volume*. En général, le CIDOC-CRM décrit des entités mais également les interactions qui peuvent exis-

ter entre ces entités : interactions entre entités persistantes ; interactions entre entités temporelles ; interactions entre entités persistantes et entités temporelles ; interactions générales entre entités (par exemple, les interactions qui existent entre entités persistantes ou temporelles avec des entités qui décrivent des durées, des lieux ou des dimensions). Il existe aussi des interactions entre entités et valeurs primitives (chaîne de caractères, nombre, date heure).

D'autre part, plusieurs projets dans le monde s'intéressent à la migration de données vers des triplestores (CIDOC-CRM ou non) : (1) Le *British Museum* (cf. http://collection.britishmuseum.org/) qui est un musée sur l'histoire et la culture et qui utilise le CIDOC-CRM ; (2) *Arches* (cf. http://www.getty.edu/conservation/our_projects/field_projects/arches/) qui est une collaboration entre le Getty Conservation Institute (GCI) et le World Monuments Fund (WMF) sur l'héritage culturel immobilier (monuments, ponts) et qui utilise le CIDOC-CRM ; (3) *DBPedia* (cf. http://www.dbpedia.org/sparql) qui est une encyclopédie en ligne largement utilisée (Ruan *et al.*, 2016) et qui utilise des langages de métadonnées comme : *dbpedia, foaf, umbel, schema.org, dublin core, geo* ; (4) *Nakala* (cf. http://www.nakala.fr/sparql) qui est un service en ligne pour déposer, documenter et diffuser des données (muséographiques) et qui utilise des langages de métadonnées comme : *foaf, skos, dublin core, vcard*.

La spécificité de notre application web est qu'elle traite de sources de données hétérogènes tant sur le contenu que sur la structure logique initiale (bases de données, fichiers XML) de ces données. Les données migrées dans des triplestores sont totalement ouvertes via notre application web. Cette application permet de visualiser les triplestores sous trois différents formats : *rdf, triplets, résumé attribut-valeur*. L'application permet aussi d'interroger les triplestores via des *"Endpoint Sparql"* (interface de saisie et d'exécution de requêtes Sparql - Sparql étant le langage de référence actuel d'interrogation de fichiers RDF) et via des *"formulaires généraux"* qui s'avèrent être utiles si on ne connait pas le Sparql (Haase *et al.*, 2004) et le CIDOC-CRM.

3. Architecture du projet ModRef

L'architecture du projet ModRef, illustrée dans la figure 2, décrit les différents processus de numérisation des données depuis la phase de création de ces données numériques à partir des connaissances d'un expert, jusqu'à son interrogation et sa visualisation par un usager. Les données peuvent ainsi subir de nombreuses transformations avant d'être finalement exploitables via des triplestores.

Ainsi, on peut passer de données non structurées ou semi structurées (notes, rapports, livres, sites web) vers des données structurées décrites par une structure logique. Cette structure logique peut être une structure à plat sous la forme *attribut-valeur* ou *tableur*, mais elle peut aussi être plus fortement structurée sous la forme de *bases de données relationnelles* ou de *fichiers XML* qui, dans notre contexte, sont des fichiers XML-EAD (Encoded Archival Description). Ces différentes descriptions font généralement usage de thésaurus (vocabulaire contrôlé de termes descripteurs ou non). À partir de ces descriptions structurelles, on va construire une description sémantique

des données sous forme de graphe sémantique RDF en s'appuyant sur des référentiels standards ou normes. Dans notre contexte, nous avons utilisé la norme CIDOC-CRM pour générer nos triplestores à partir d'un alignement de nos données avec le graphe sémantique CIDOC-CRM. Ces triplestores vont pouvoir être exploités dans diverses applications web sémantique ou via des "Endpoint Sparql".

La première phase de transformation des données (données non structurées ou semi structurées vers données structurées) est réalisée au sein de chaque sous projet tandis que le projet ModRef lui intervient principalement dans la deuxième phase de transformation des données (données structurées vers données sémantiques).

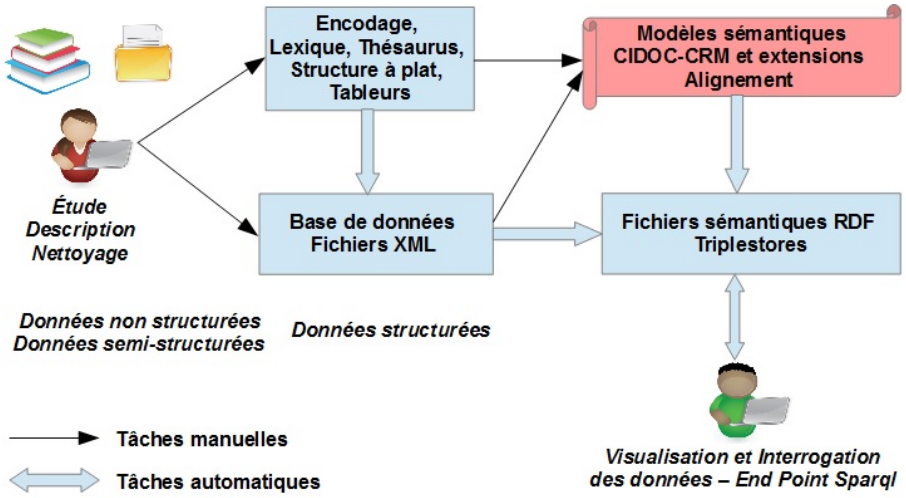


Figure 2. Architecture du projet ModRef.

Par ailleurs, pour réaliser la preuve conceptuelle de ModRef, trois projets pilotes ont été sélectionnés :

1. *CDLI* : conservatoire numérique ou musée virtuel de l'ensemble des documents rédigés en écriture cunéiforme (cf. http://www.cdli.ucla.edu) ;

2. *ObjMythArcheo* : corpus numérique d'objets archéologiques à iconographie mythologique (cf. http://www.limc-france.fr et http://medaillesetantiques.bnf.fr) ;

3. *BiblioNum* : bibliothèque numérique sur l'histoire de France du 20e siècle (cf. http://www.argonnaute-u.paris10.fr).

La table 1 compare les données des trois projets pilotes de ModRef sur 5 critères : taille des textes, nombre d'objets, type de la structure logique, nombre d'éléments de la structure logique et langue de description des données.

4. Modélisation CIDOC-CRM de ModRef et alignement des données

Nous avons identifié les classes CIDOC-CRM utiles (dont au moins un chemin conduit vers une valeur non nulle) pour modéliser les données de nos trois projets pi-

Table 1. Comparaison des données des projets pilotes de ModRef.

	CDLI	**ObjMythArcheo**	**BiblioNum**
Taille textes	300 Mo	100 Mo	100 Mo
Nombre d'objets	313 332 tablettes	17 424 objets	77 collections - 62 392 fichiers
Structure logique	Base de données de type Tableur	Base de données relationnelle	XML-EAD
Nombre d'éléments de structure	1 table avec 61 attributs	59 tables	146 éléments XML-EAD
Langue	Anglais	Français-Anglais	Français

lotes. Cela représente des extraits de graphes relatifs aux quatre thèmes suivants : (1) caractéristiques générales (identifiant, type, titre, matériau, dimension, note/description), bibliographie, composition et similarité d'objets ; (2) évènements de début d'existence (origine) et de fin d'existence ; (3) activités diverses (transfert de garde, mesure, conservation) ; (4) inscriptions et autres représentations (photo, dessin, tableau).

De façon générale, ces extraits sont stables pour tout projet car dans le CIDOC-CRM, il est possible d'identifier tous les chemins possibles pour obtenir une information donnée sur un objet. En effet, un graphe sémantique est un ensemble de nœuds et d'arcs orientés ou relations qui obéissent à un certain nombre de contraintes et règles (raccourci, héritage, inverse, symétrie, transitivité). Ce sont ces contraintes et règles qui définissent la cohérence et la validité d'un modèle.

Dans cette section, nous décrirons nos différents thèmes de modélisation de graphe CIDOC-CRM ainsi qu'un exemple d'alignement. En effet, le principe d'alignement est globalement le même pour tous les thèmes et pour tous les projets pilotes.

4.1. Modélisation des caractéristiques générales

Les caractéristiques générales d'un objet s'obtiennent le plus souvent par des interactions, avec des chemins de graphe assez courts, entre entités. Elles permettent de définir pour un objet les éléments suivants : identifiant, type (catégorisation), titre, matériau, dimension, note.

La modélisation des caractéristiques générales des objets du projet ModRef est illustrée dans la figure 3. Dans cette figure, on peut observer l'existence de deux chemins de graphes différents pour la définition des dimensions d'un objet : (1) un *chemin plus court* ou *raccourci* qui relie la classe *E70 Thing* à la classe *E54 Dimension* avec la propriété *P43 has dimension*, soit le triplet *[E70 Thing, P43 has dimension, E54 Dimension]* ; (2) un *chemin plus long* qui contient davantage de nœuds informationnels à remplir. Ce chemin est décrit par les triplets suivants : *[E1 CRM Entity, P39i was measured by, E16 Measurement], [E16 Measurement, P40 observed dimension, E54 Dimension]*. Avec ce chemin, on peut en plus remplir des informations concernant l'activité de mesure *E16 Measurement*. En effet, la classe *E16 Measurement* est

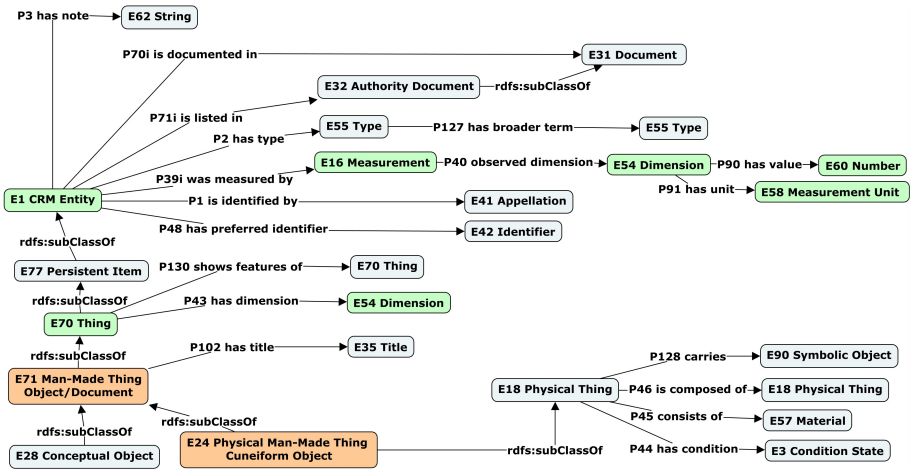


Figure 3. Modélisation des caractéristiques générales.

un type d'activité car les classes *E13 Attribute Assignment*, *E7 Activity* et *E5 Event* font partie de sa hiérarchie (cf. figure 4).

Il est tout à fait possible de remplir différents chemins donnant une même information dans un graphe. Cependant, on peut être amené à faire un choix entre deux possibilités de chemins lorsque l'on ne dispose pas des informations nécessaires pour décrire un chemin donné. Ceci est le cas le plus souvent lorsqu'une entité temporelle fait partie du chemin. D'autre part, la figure 3 permet aussi d'illustrer d'autres interactions entre entités persistantes comme : *P70i is documented in* pour les références bibliographiques, *P46 is composed of* pour la composition d'objets, *P130 shows features of* pour la similarité entre objets, *P128 carries* pour la relation entre un objet et une entité qui se trouve sur l'objet, comme une inscription par exemple.

4.2. Modélisation des évènements de début et de fin d'existence

Une activité importante concernant les informations muséographiques consiste à décrire leur origine : à les dater, à définir leur lieu d'origine et éventuellement les participants à leur création. La modélisation des évènements de début et de fin d'existence des objets du projet ModRef est illustrée dans la figure 4. Le CIDOC-CRM permet ainsi de définir la date, le lieu et les participants de chaque évènement.

Pour le début d'existence (origine), on utilise l'évènement *E63 Beginning of Existence* et les patterns de triplets suivants : *[E77 Persistent Item, P92i was brought into existence by, E63 Beginning of Existence], [E2 Temporal Entity, P4 has time-span, E52 Time-Span], [E52 Time-Span, P78 is identified by, E49 Time Appellation], [E4 Period, P7 took place at, E53 Place], [E5 Event, P11 had participant, E39 Actor], [E63 Beginning of Existence, rdfs : subClassOf, E5 Event], [E5 Event, rdfs : subClassOf, E4 Period], [E4 Period, rdfs : subClassOf, E2 Temporal Entity]*. Par ailleurs, on

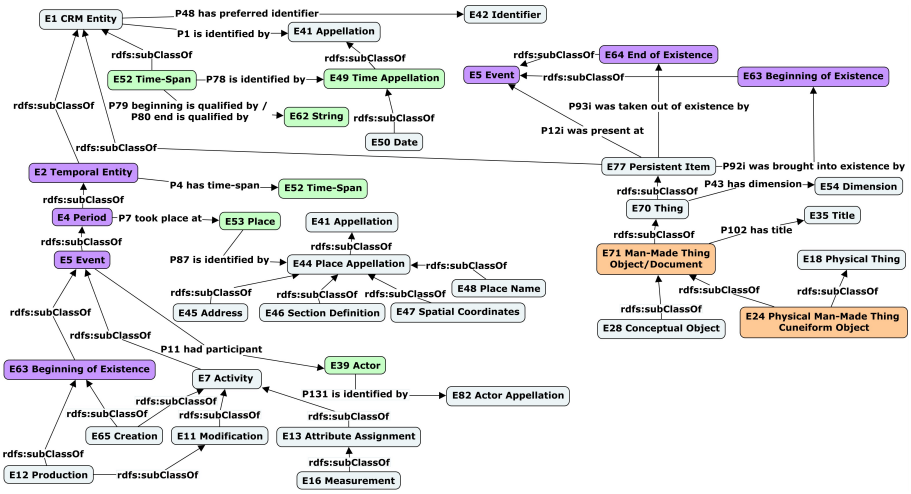


Figure 4. Modélisation des évènements de début et de fin d'existence.

peut aussi partir des activités *E65 Creation* ou *E12 Production* qui ont pour superclasses les classes *E63 Beginning of Existence* et *E7 Activity* (cf. figure 4).

Pour la fin d'existence, on utilise la classe *E64 End of Existence* ou une de ses sous-classes. Ainsi, on va pouvoir définir également la date, le lieu et les participants à la fin d'existence d'un objet.

4.3. Modélisation des activités

La figure 5 illustre un extrait de notre modèle pour la description des activités en général, et de l'activité de *transfert de garde* en particulier. Ainsi, pour rattacher un objet à une activité de transfert de garde, on va utiliser la propriété *P30 transferred custody of* (ou son inverse *P30i custody transferred through*) entre l'activité (*E10 Transfer of Custody*) et l'objet physique (*E18 Physical Thing*). De plus, pour un transfert de garde, on peut décrire les différents protagonistes du transfert (*P29 custody received by, P28 custody surrendered by*) et décrire éventuellement aussi un historique des différents transferts de garde relatifs à un objet ou à un document donné. Notons qu'il existe également un chemin de raccourci qui ne passe pas par l'activité de transfert de garde et qui permet de définir les gardiens ou propriétaires anciens ou actuels d'un objet (*P49 has former or current keeper, P50 has current keeper, P51 has former or current owner, P52 has current owner*).

De façon générale, pour un évènement ou une activité, on va décrire la date, le lieu et les participants ou acteurs de l'évènement ou de l'activité. Plus spécifiquement pour une activité (transfert de garde, assignation de valeur à un attribut, mesure, conservation), on va pouvoir en plus décrire : la procédure utilisée (*P33 used specific technique, P32 used general technique*), les objets utilisés (*P16 used specific object,*

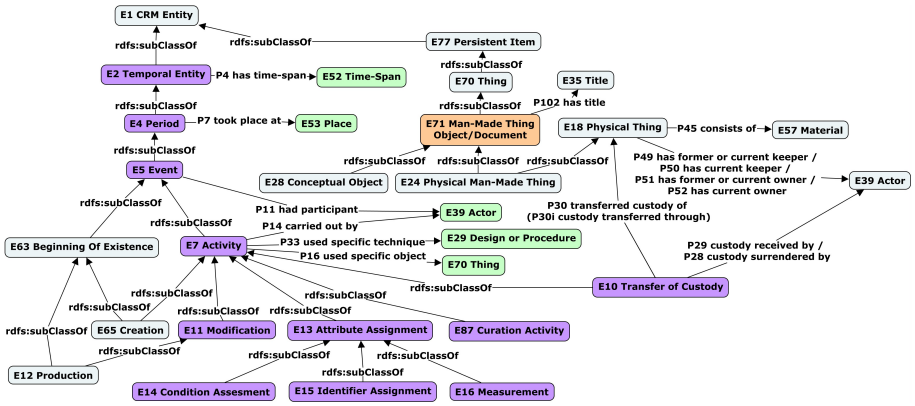


Figure 5. Modélisation des activités.

P125 used object of type), l'objectif de l'activité (*P20 had specific purpose, P21 had general purpose*).

4.4. Modélisation des inscriptions et autres représentations d'un objet

La figure 6 illustre un extrait de notre modèle pour la description des inscriptions sur des objets ou la description d'autres représentations (photos, dessins, tableaux) de ces objets.

Ainsi, pour rattacher un objet à son inscription, on va utiliser la propriété *P128 carries* entre un objet physique (*E18 Physical Thing*) et un objet symbolique (*E90 Symbolic Object*) qui se trouve sur l'objet et qui est ici notre inscription. Cela permet donc de retrouver, par exemple, les objets qui portent une certaine inscription (sceau, signature).

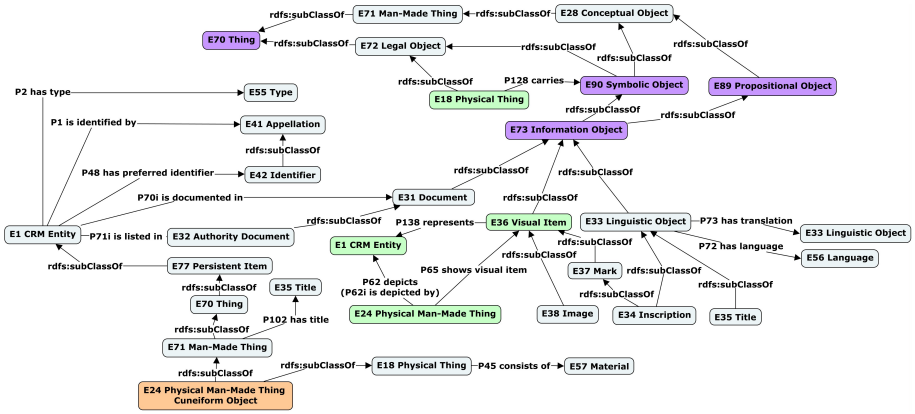


Figure 6. Modélisation des inscriptions et autres représentations d'un objet.

Par ailleurs, pour relier une photo (ou dessin ou tableau) à un objet (physique ou conceptuel), on peut utiliser un ensemble de propriétés : (1) *P62 depicts* pour décrire le lien entre la photo ou le dessin ou le tableau (ici, *E24 Physical Man-Made Thing*) et l'entité *E1 CRM Entity* (objet physique ou conceptuel) *représentée* par la photo ou le dessin ou le tableau. Cette propriété ne se rapporte pas aux inscriptions ou autres informations encodées sur un objet ; (2) *P65 shows visual item* permet de rattacher la photo ou le dessin ou le tableau à une représentation visuelle (*E36 Visual item*) de l'objet *représenté* par la photo ou le dessin ou le tableau ; (3) *P138 represents* permet de relier une représentation visuelle (*E36 Visual item*) d'un objet à l'objet en question (*E1 CRM Entity*). Notons cependant que la propriété *P62 depicts* est un raccourci des propriétés *P65 shows visual item et P138 represents*. La photo ou dessin ou tableau étant généralement décrite avec la classe *E24 Physical Man-Made Thing*.

4.5. Alignement des données

La migration de données vers des triplestores nécessite une phase d'alignement des données avec les extraits de graphe sémantique CIDOC-CRM proposés. Cet alignement est indispensable du fait de l'hétérogénéité initiale de la description des données, conséquence aussi de la diversité des projets pilotes de ModRef. Cet alignement n'est pas une tâche programmatique mais fait appel à des détails de structure logique propre au modèle de description de données choisi par chaque projet. C'est une tâche à mi-chemin entre la modélisation et l'implémentation qu'elle permet d'entrevoir un peu plus clairement. Notons que cette tâche ne doit pas être confondue avec l'alignement entre ontologies (fichiers owl/rdf) (Faria *et al.*, 2013) étant donné que notre alignement est plutôt entre l'ontologie du CIDOC-CRM et des données brutes provenant de bases de données ou de fichiers XML (dans notre contexte, des fichiers XML-EAD).

L'alignement va consister principalement à remplir les nœuds du graphe sémantique. Les nœuds terminaux vont être remplis par des valeurs extraites des structures logiques des données des projets correspondants et les nœuds non-terminaux ou intermédiaires seront remplis avec des URIs qui définissent ainsi des chemins vers les nœuds terminaux. Notons qu'une rigueur particulière doit être apportée à la construction des URIs, à la fois pour leur lisibilité mais également pour la cohérence des chemins dans le graphe, afin d'éviter des conflits de chemins et garantir ainsi l'unicité d'un chemin donné par rapport à un autre. La figure 7 illustre un extrait d'alignement de données, initialement au format XML-EAD (cf. https://www.loc.gov/ead/tglib/element_index.html) et correspondant au premier thème de notre modélisation sémantique (cf. figure 3). En XML-EAD, pour obtenir par exemple les dimensions d'un objet on utilise les chemins d'accès xpath "*/ead/archdesc/did/physdesc/dimensions*" ou *"/ead/archdesc/dsc/c/did/physdesc/dimensions*", selon que l'on est au niveau de la collection ou du document. Ainsi, pour décrire les dimensions d'un objet, on utilise une succession de triplets de la forme :

[http ://www.modref.org/biblionum/document_id/e70_thing, rdf : type, ecrm : E70_Thing],

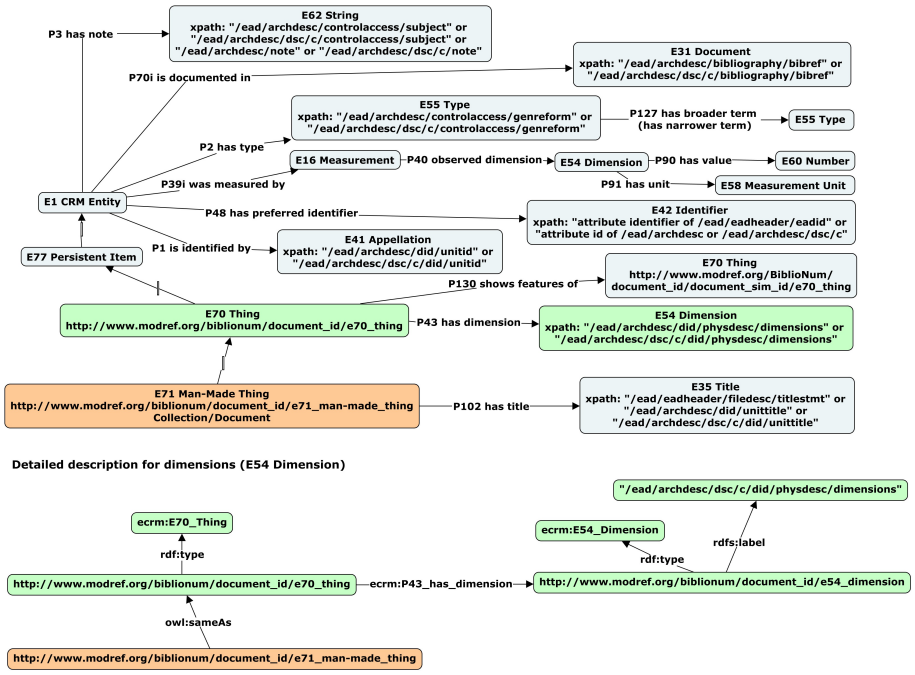


Figure 7. Exemple d'alignement de données au format XML-EAD.

[http ://www.modref.org/biblionum/document_id/e70_thing, ecrm : P43_has_dimension, http ://www.modref.org/biblionum/document_id/e54_dimension],

[http ://www.modref.org/biblionum/document_id/e54_dimension, rdf : type, ecrm : E54_Dimension],

[http ://www.modref.org/biblionum/document_id/e54_dimension, rdfs : label, "/ead/archdesc/dsc/c/did/physdesc/dimensions"],

[http ://www.modref.org/biblionum/document_id/e71_man-made_thing, owl : sameAs, http ://www.modref.org/biblionum/document_id/e70_thing].

De façon générale, l'alignement réalisé va être décrit dans une structure de données programmatique qui sera utilisée pour générer automatiquement des fichiers qui respectent la syntaxe RDF et CIDOC-CRM : c'est la migration des données ou création de nos triplestores.

5. Migration de données vers des triplestores

Une migration efficace et cohérente de données fait appel à différentes compétences. Pour assurer la pérennisation de cette procédure, une architecture générale et rigoureuse du workflow des différents types de données à manipuler doit être définie.

Cette architecture explicite la démarche globale de tout projet qui souhaite faire migrer ses données vers des triplestores. Pour ModRef, cette démarche se subdivise en différentes étapes bien identifiées : préparation des données (étude et description structurelle), modélisation sémantique et alignement des données structurées avec le graphe sémantique CIDOC-CRM, création et exposition de triplestores qui vont alors pouvoir être interrogés par des usagers ou des applications web sémantique. Initialement, les données sont souvent non-structurées ou semi-structurées (notes, rapports, livres, html) et ont besoin d'être d'abord décrites par une représentation structurée (tableurs, bases de données, fichiers XML-EAD) afin de pouvoir construire plus facilement leur représentation sémantique, par la suite. Ce continuum d'étapes fait intervenir plusieurs sous-procédures pour assurer le passage d'un format de représentation de données à un autre. Ainsi, on distingue deux phases principales permettant le passage : (1) des données non structurées ou semi structurées vers des données structurées ; (2) des données structurées vers des données sémantiques.

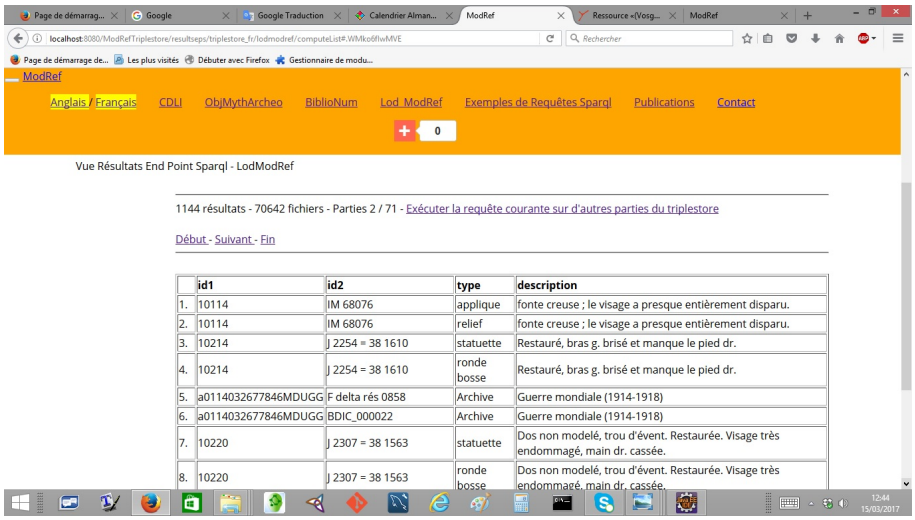
En effet, l'élément clé du processus de migration de données vers des triplestores est la modélisation et l'alignement des données avec le modèle de graphe sémantique choisi. Pour réaliser un alignement des données avec notre graphe CIDOC-CRM, nous avons effectué une mise en correspondance de certains nœuds du graphe sémantique proposé avec des informations extraites à la fois de bases de données mais aussi de collections de fichiers XML-EAD. Cette migration implique donc à la fois de la lecture de bases de données et du parsing de fichiers XML-EAD (voir la table 1).

La preuve conceptuelle du projet ModRef ou la validation de la migration de données vers des triplestores concernent donc un ensemble de tâches en amont (préparation et structuration des données, modélisation sémantique, alignement des données) et en aval (exposition, visualisation, interrogation et exploration des données) du processus de migration. Ainsi, l'exploitation des triplestores créés et les bénéfices que l'on peut en tirer est l'autre aspect majeur autour de la question de ces nouveaux entrepôts de documents RDF que sont les triplestores.

6. Visualisation et Exploitation de triplestores

Les triplestores créés sont exposés pour consultation (sous trois formes : *rdf, triplets, résumé attribut-valeur*) mais aussi pour interrogation via notre application web. L'intérêt du triplestore est qu'on a un modèle connu public et publié de représentation de l'information, ce qui permet d'interroger les triplestores indifféremment avec des procédures identiques. Nous avons défini deux procédures d'exploitation de nos triplestores : des interfaces sous forme de *"formulaires généraux"* et des *"Endpoint Sparql"* (cf. figure 8).

Les formulaires sont un moyen simple et assez intuitif, car très proche du langage naturel, pour formuler des requêtes vers nos triplestores. Nul besoin donc de compétences particulières, il suffit de remplir les rubriques du formulaire qui nous intéressent et de lancer la recherche. Une requête Sparql est automatiquement construite à partir des valeurs des champs renseignés du formulaire et c'est cette requête qui est utilisée



	id1	id2	type	description
1.	10114	IM 68076	applique	fonte creuse ; le visage a presque entièrement disparu.
2.	10114	IM 68076	relief	fonte creuse ; le visage a presque entièrement disparu.
3.	10214	J 2254 = 38 1610	statuette	Restauré, bras g. brisé et manque le pied dr.
4.	10214	J 2254 = 38 1610	ronde bosse	Restauré, bras g. brisé et manque le pied dr.
5.	a0114032677846MDUGG	F delta rés 0858	Archive	Guerre mondiale (1914-1918)
6.	a0114032677846MDUGG	BDIC_000022	Archive	Guerre mondiale (1914-1918)
7.	10220	J 2307 = 38 1563	statuette	Dos non modelé, trou d'évent. Restaurée. Visage très endommagé, main dr. cassée.
8.	10220	J 2307 = 38 1563	ronde bosse	Dos non modelé, trou d'évent. Restaurée. Visage très endommagé, main dr. cassée.

Figure 8. Application Web du projet ModRef : Endpoint Sparql.

pour interroger le triplestore. Au terme de l'exécution de la requête, une liste d'objets sélectionnés est renvoyée en résultat à l'usager qui peut les consulter également sous trois formes : *rdf, triplets, résumé attribut-valeur*. Par ailleurs, on peut aussi interroger nos triplestores via des "Endpoint Sparql". Ce deuxième mode d'interrogation nécessite la connaissance du langage Sparql qui est aujourd'hui le langage de référence pour l'interrogation de documents RDF. Sparql est un langage assez simple mais pas toujours à la portée de tous. Ainsi, les formulaires généraux peuvent être vus comme un premier point d'entrée pour l'interrogation des triplestores tandis que les "Endpoint Sparql" assurent une exploitation (interrogation et exploration) plus large de ces derniers via une formulation libre de requêtes Sparql de type "Select".

Notre application web permet de consulter, d'interroger et d'explorer nos triplestores séparément pour chaque projet pilote de ModRef mais aussi en regroupant les triplestores via le LOD (Linked Open Data) de ModRef. L'application web offre, pour chaque projet et pour le LOD de ModRef, la possibilité de consulter les données sous trois formes mais aussi celle de les interroger via des "formulaires généraux" mais aussi via des "Endpoint Sparql". Ainsi, en résultat d'une requête, le LOD permet de retrouver des informations diverses (statue/statuette, archive) provenant de différents triplestores (cf. figure 8). Plusieurs requêtes Sparql ont été exécutées pour valider la migration de données et une liste de requêtes exemples est fournie dans notre application web. Nous avons développé notre propre "Endpoint Sparql" et nous offrons également la possibilité d'interroger nos données via un "Endpoint Sparql" Virtuoso (logiciel permettant de créer un lien internet vers une instance de "Endpoint Sparql") disponible via le lien suivant : http://3s-passespresent.huma-num.fr/sparql.

Notons que la notion d'exploitation de triplestores fait appel aux notions d'interrogation et d'exploration de graphe. Ainsi, l'interrogation de triplestores consiste à

formuler une requête Sparql pré-formatée (formulaires généraux) ou libre (Endpoint Sparql) tandis que l'exploration de triplestores est une forme d'interrogation uniquement possible via des "Endpoint Sparql" qui permet de découvrir différents chemins dans un graphe sémantique vers des données précises. En effet, plusieurs chemins peuvent permettre d'obtenir une même information dans un graphe (usage de diverses notions : raccourci, raffinement, héritage, inverse) sachant que ces chemins ne sont pas toujours tous renseignés. On peut donc écrire des requêtes Sparql pour découvrir si différents chemins vers une donnée précise existent ou pour connaitre les chemins menant vers des nœuds terminaux. L'exploration est donc importante pour s'approprier un triplestore CIDOC-CRM spécifique.

7. Procédure d'évaluation et résultats

Table 2. Requêtes Sparl.

(a) Liste des triplets terminaux

```
SELECT Distinct ?subject ?predicate ?object
WHERE {
?subject ?predicate ?object .
Filter ( isLiteral( ?object) && ?object != "" )
}
```

(b) Liste des types d'objets

```
PREFIX ecrm : <...>
PREFIX rdf : <...>
PREFIX rdfs : <...>
SELECT Distinct ?type
WHERE {
 ?type_uri rdf : type ecrm : E55_Type .
 ?type_uri rdfs : label ?type .
Filter ( ?type != "" ) }
```

(c) Liste de caractéristiques provenant de l'entité "E1 CRM Entity".

```
PREFIX ecrm : <http ://erlangen-crm.org/150929/>
PREFIX rdf : <http ://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs : <http ://www.w3.org/2000/01/rdf-schema#>
SELECT Distinct ?id1 ?id2 ?type ?description
WHERE {
 ?e1_obj ecrm : P48_has_preferred_identifier ?id1_uri .
 ?id1_uri rdfs : label ?id1 .
 ?e1_obj ecrm : P1_is_identified_by ?id2_uri .
 ?id2_uri rdfs : label ?id2 .
 ?e1_obj ecrm : P2_has_type ?type_uri .
 ?type_uri rdfs : label ?type .
 ?e1_obj ecrm : P3_has_note ?description . }
```

Nous avons conçu et exécuté plusieurs requêtes Sparql pour valider les différents datasets de nos triplestores. Les requêtes sont divisées en deux groupes, un premier groupe relatif au schéma de la syntaxe RDF (*liste des concepts ou des prédicats utilisés, liste des triplets terminaux (cf. Table 2a), liste des triplets d'une ressource donnée, extraits de chemins menant vers des nœuds terminaux non vides*) et un autre groupe relatif au schéma de la norme CIDOC-CRM (*vérification de l'instanciation d'une classe spécifique, vérification des labels d'une entité ou ressource donnée (cf. Table 2b), ca-*

ractéristiques générales d'un objet (cf. Table 2c), information sur l'origine ou la garde d'un objet).

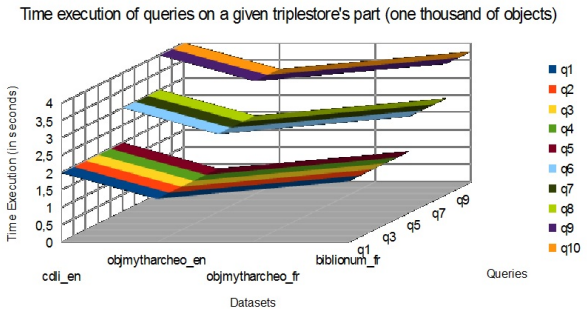


Figure 9. Exécution de requêtes Sparql pour le projet ModRef.

De plus, les triplestores sont subdivisés en parties constantes (nombre d'objets ou nombre de triplets) et les requêtes sont exécutées chaque fois sur une seule partie et puis progressivement sur les autres parties si l'utilisateur le demande. Les résultats sont donc fusionnés au fur et à mesure. L'utilisateur peut arrêter l'exécution sans avoir à exploiter tout le triplestore. Le numéro de la partie courante (sur laquelle vient de s'exécuter la requête courante) et le nombre total de parties du triplestore sont toujours affichés. La figure 9 montre que le temps moyen d'exécution (en secondes) des requêtes sur une partie de triplestore (*soit 1000 objets, approximativement 100 000 triplets*) est plutôt constant et la rapidité d'exécution de ces requêtes est tout à fait acceptable pour les usagers. Par contre, le temps d'exécution cumulatif augmente si l'on couvre davantage de parties du triplestore.

8. Conclusion

Le projet ModRef permet de réaliser une preuve conceptuelle de la migration de données vers des triplestores CIDOC-CRM à travers : une architecture générale qui identifie les différentes étapes à suivre ; la modélisation et l'alignement des données avec le graphe sémantique ; la migration des données vers les triplestores ; l'exposition des triplestores via l'application web bilingue "anglais-français" http://triplestore .modyco.fr qui permet de consulter, d'interroger et d'explorer ces triplestores.

Les perspectives qui découlent de nos travaux concernent : (1) *le partage, l'échange et la découverte de connaissances à plus grande échelle* en intégrant d'autres LOD (Linked Open data) sur internet (Beek, Rietveld *et al.*, 2016) (Daga *et al.*, 2016). Le LOD doit améliorer la découverte de nouvelles connaissances, du fait de la quantité et de la diversité des données liées mais surtout du fait de l'usage de formalismes, de langages de métadonnées, de thésaurus publiés, standardisés voire normalisés ; (2) *la comparaison de graphes sémantiques* qui décrivent des données similaires (Beek, Schlobach, Harmelen, 2016) (objets ressemblants, objets d'une même période histo-

rique, objets de même type, objets identiques) dans un contexte de LOD. Il en résultera un enrichissement mutuel des différents acteurs ou usagers des LOD.

Acknowledgements

L'auteur remercie le laboratoire Labex "Les passés dans le présent" de l'Université de Paris 10 et le projet ANR ModRef de référence ANR-11-LABX-0026-01.

Références

Beek W., Rietveld L., Schlobach S., Harmelen F. van. (2016). Lod laundromat : Why the semantic web needs centralization (even if we don't like it). *IEEE Internet Computing*, Vol. 20, N° 2, pp. 78-81.

Beek W., Schlobach S., Harmelen F. van. (2016). A contextualised semantics of owl :sameas. *In proceedings of the 13th Extended Semantic Web Conference ESWC'16*, pp. 405-419.

Berners-Lee T., Hendler J., Lassila O. (2001). The semantic web. *Scientific American*, pp. 34-43.

Boeuf P. L., Doerr M., Ore C. E., Stead S. (2015, May). Definition of the cidoc conceptual reference model, version 6.2. *Produced by the ICOM/CIDOC Documentation Standards Group, Continued by the CIDOC CRM Special Interest Group. http ://www.cidoc-crm.org/ [retrieved : March, 2017].*

Daga E., d'Aquin M., Adamou A., Brown S. (2016). The open university linked data - data.open.ac.uk. semantic web. *Semantic Web*, Vol. 7, N° 2, pp. 183-191.

Faria D., Pesquita C., Santos E., Palmonari M., Cruz I., Couto F. (2013). The agreementmakerlight ontology matching system. *The 12th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE)*, pp. 527-541.

Haase P., Broekstra J., Eberhart A., Volz R. (2004). Comparison of rdf query languages. *In proceedings of the third International Semantic Web Conference ISWC'04*, pp. 502-517.

Hooland S. V., Verborgh R. (2014). *Linked data for libraries, archives and museums. how to clean, link and publish your matadata* (A. L. Association, Ed.).

Oldman D., Doerr M., Jong G. de, Norton B., Wikman T. (2014). Realizing lessons of the last 20 years : A manifesto for data provisioning and aggregation services for the digital humanities. http ://www.dlib.org/dlib/july14/oldman/07oldman.html [retrieved : March, 2017]. *D-Lib Magazine*, Vol. 20, N° 7/8.

Ruan T., Li Y., Wang H., Zhao L. (2016). From queriability to informativity, assessing "quality in use" of dbpedia and yago. *In proceedings of the 13th Extended Semantic Web Conference ESWC'16*, pp. 52-68.

Scarinci J., Myers T. (2014). A semantic web framework to enable sustainable lodging best management practices in the usa. *Information Technology and Tourism*, Vol. 14, N° 4, pp. 291-315.

Shadbolt N., Berners-Lee T., Hall W. (2006). The semantic web revisited. *IEEE Intelligent Systems*, Vol. 21, N° 3, pp. 96-101.

Proposition d'une démarche de construction d'une cartographie des connaissances

Sahar GHRAB[1, 2], Inès SAAD[2, 3], Gilles KASSEL[2], Faiez GARGOURI[1]

1. Laboratoire MIRACL, Université de Sfax
ISIMS, Pôle technologique de Sfax, Sakiet Ezzit, 3021Sfax, Tunisie
ghrab.sahar@gmail.com

2.Laboratoire MIS, Université de Picardie Jules Verne
14 Quai de la Somme, 80080 Amiens, France

3.Amiens Business School
18 Place Saint-Michel, 80000 Amiens, France

RESUME. Le papier présente un système d'informations et de connaissances baptisé CK-Cartography. Son objectif est l'aide à la cartographie des savoir-faire et des savoirs factuels via une visualisation intuitive, compréhensible et conviviale permettant de partager un langage commun de communication entre les membres de l'organisation. Afin de mieux concevoir CK-Cartography, nous proposons un méta-modèle spécifique à la cartographie des savoirs. Ce méta-modèle propose les différents concepts à cartographier ainsi que les relations qui les interconnectent. Il est basé sur une approche ontologique permettant de définir les concepts. CK-Cartography utilise un langage graphique pour la représentation visuelle de ces concepts. Il est expérimenté et validé dans l'Association de Sauvegarde des Handicapés Moteurs de Sfax (ASHMS).

ABSTRACT. The paper presents information and knowledge system named CK-Cartography. Its goal is to help Know-How and Knowing-That cartography through intuitive, understandable and user-friendly visualization allowing to share a communication common language between members of organization. In order to well design CK-Cartography, we propose a meta-model for knowledge cartography. This meta-model proposes the different concepts to be mapped as well as the relations that interconnect them. It is based on an ontological approach for concepts' rigorous definitions. CK-Cartography uses a graphical language for concepts' visual representation. It is experimented and validated in the Association of Protection of Motor Disabled of Sfax (ASHMS).

Mots-clés :cartographie des connaissances, savoir-faire, savoir factuel, CK-Cartography, méta-modèle de cartographie des savoirs, langage graphique, approche de cartographie des savoirs

KEYWORDS: Knowledge Cartography, Know-How, Knowing-That, CK-Cartography, knowledge cartography meta-model, graphical language, knowledge cartography approach

1. Introduction

A l'ère du numérique, les technologies de l'information et de la communication implémentées dans les systèmes d'information ne se limitent pas à échanger les informations entre les individus mais à s'assurer du sens qu'elles véhiculent, des interprétations qu'elles vont prendre et des connaissances créées qu'elles en résultent (Arduin et al., 2015; Saad et al., 2017). Ces connaissances peuvent être soit des savoir-faire soit des savoirs factuels. Les savoir-faire sont souvent tacites incarnés dans la tête de leurs détenteurs et les savoirs factuels sont stockés dans des supports qui peuvent être numériques ou sous forme papier (Ghrab et al., 2016). En effet, le système d'information et de connaissance ne véhicule pas simplement des informations mais contient aussi des savoir-faire et des savoirs factuels dont les individus sont porteurs et ayant des schémas d'interprétation différents. L'interprétation des informations par les individus peut créer les savoir-faire (tacites ou explicités) et les savoirs factuels. Le système proposé consiste à mettre en œuvre les individus détenant les connaissances afin de mieux les repérer et les valoriser. Plusieurs travaux dans la littérature montrent l'intérêt de la cartographie des connaissances dans les organisations.

Dans cet article, nous proposons un système inter-organisationnel CK-Cartography de cartographie des savoir-faire et des savoirs factuels. Son objectif est l'identification et la visualisation des savoir-faire et des savoirs factuels. L'expérimentation de CK-Cartography est menée dans le domaine médical au profit de l'Association de sauvegarde des handicapés Moteurs de Sfax (ASHMS). L'enjeu envisagé est de mieux identifier les savoir-faire et les savoirs factuels de l'organisation et de faciliter leur partage entre les professionnels de santé affiliés à l'ASHMS ou à d'autres organisations. Cet enjeu est assuré par l'outil CK-Cartography implémenté dans l'ASHMS et utilisé par différents professionnels de santé (les médecins bénévoles, les professeurs des universités, les techniciens de santé, etc.).

Le plan de cet article est structuré comme suit. Dans la deuxième section, nous détaillerons les travaux de cartographie des connaissances. Dans la troisième section, nous proposerons notre méta-modèle de cartographie des savoir-faire et des savoirs factuels. Dans la quatrième section, nous détaillerons le langage graphique iconique proposé. Les résultats d'expérimentation de CK-Cartography dans l'ASHMS feront l'objectif de la section suivante. Nous énumérons, dans une autre section, la démarche adoptée pour la construction de CK-Cartography. Finalement, nous rappellerons les contributions apportées dans cet article et nous envisagerons quelques perspectives d'ouverture.

2. La cartographie des connaissances: état de l'art

Plusieurs travaux, dans la littérature, ont été proposés ayant pour objectif la cartographie des connaissances. Nous citons ceux de (Wickel et al., 2013; Bresciani et Eppler, 2013; Hao et al., 2014; Dorze et al., 2014; Balaid et al., 2013). Chaque cartographie des connaissances a un objectif bien défini (Pourquoi ?), des concepts spécifiques à cartographier (Quoi ?), des techniques de visualisation (Comment ?) et une approche de cartographie à adopter pour la visualisation des connaissances.

Dans cette section, nous décrivons quelques travaux en cartographie des connaissances: (Tricot, 2006 ; Crampes et al., 2008 ;Gandon, 2008 ; Sellin, 2011).

Sellin (2011) propose une cartographie des connaissances pour le pilotage des ressources humaines et le transfert des connaissances entre les membres de l'organisation. L'individu dans son poste se situe au centre de la carte de connaissances qui contient les branches de «clients, de «ressources», d'«activité», de «livrables» et de «savoirs clés». Cette cartographie favorise l'identification des connaissances clés, positionne l'individu dans l'organisation de façon à montrer son cadre de travail, les membres de son équipe, ce qu'il produit à son successeur et reçoit de son prédécesseur. Inspirée du modèle SECI (Socialisation, Externalisation, Combinaison, Internalisation), Sellin (2011) propose un codage de couleur pour le transfert des savoir-faire. Une autre caractéristique de la cartographie des connaissances proposée par Sellin (2011) consiste à utiliser les nombres pour référer la durée d'usage de la connaissance dans l'organisation.

Tricot (2006) propose une cartographie sémantique de l'espace informationnel de l'organisation basée sur la sémantique du domaine traité. La cartographie sémantique est une cartographie de connaissances permettant l'échange et le partage des connaissances au sein de l'entreprise OntologosCorp et le groupe d'expertise comptable SADEC. Cette cartographie est guidée par une ontologie mettant en œuvre les concepts de poste, d'emploi, de métier et de filière. Pour la construction de la cartographie sémantique, (Tricot, 2006) propose une démarche de construction de la carte de connaissances composée de quatre phases: la structuration de l'espace informationnel brut, la représentation de l'espace informationnel structuré, la visualisation de la carte représentée et l'adaptation de la carte par l'interaction de l'utilisateur avec les espaces de cartographie.

Crampes et al. (2008) proposent une cartographie sémantique auto-organisée d'un référentiel de connaissances partagé. Cette cartographie permet la visualisation du référentiel à travers la représentation de réseaux sous forme de graphes. Elle est basée sur l'environnement de dessin de graphes Molage. La représentation des connaissances du référentiel est assurée par des vues globales et locales. Chaque vue intègre les documents, les auteurs et les concepts de l'ontologie du domaine Molage. Dans la cartographie des connaissances, les auteurs sont disposés verticalement par ordre alphabétique, les documents sont présentés verticalement par rapport aux auteurs à l'aide de la relation "auteur-de". Crampes et al. (2008) utilisent des cercles pour désigner les connaissances dans la carte de connaissances ainsi que les icônes pour mettre en valeur d'autres concepts.

Gandon (2008) propose une cartographie des compétences pour la Télécom Valley proposée dans le cadre du projet KmP (*Knowledge Management Platform*) qui est construit autour de scénarios d'usage pour la visualisation, l'échange et le partage des compétences entre les entreprises et les organismes de recherche de la Télécom Valley. Les compétences à cartographier sont des compétences collectives qui sont le fruit d'une combinaison de compétences plus élémentaires, en particulier individuelles et nécessite une coopération entre les membres de l'équipe opérationnelle. Dans la cartographie, le cluster regroupe des acteurs aux compétences complémentaires (qui relèvent du même système d'offre) et un pôle regroupe des acteurs aux compétences similaires (qui réalisent le même type d'action ou utilisent les mêmes ressources). L'utilisateur peut identifier des clusters (ou des pôles) de compétences dans la Télécom Valley. Les clusters sont représentés sous forme de « grappes » et de « bulles » affichées sur une sorte de « radar ». Les grappes correspondent à un ensemble de ressources technologiques, scientifiques ou managériales. Les bulles correspondent à des actions.

Les différents travaux déjà détaillés montrent que la cartographie des connaissances est utilisée comme moyen soit pour l'identification ou pour le partage des connaissances. Les connaissances à partager sont incarnées dans la tête des individus qui les détiennent (connaissances tacites) ou stockées souvent dans des bases de données, des bases de connaissances ou des documents numériques (connaissances explicitées). Les cartographies des connaissances proposées, malgré leur ergonomie, ne montrent pas visuellement les relations qui existent entre les différents nœuds représentant les concepts de connaissance, de support, d'acteur, d'action et etc. Il est intéressant de visualiser ces concepts et les liens qui les relient pour distinguer d'une part les connaissances tacites des connaissances explicitées et d'autre part les savoir-faire qui sont liés aux actions des savoirs factuels qui sont liés aux descriptions.

3. Méta-modèle de cartographie des savoirs

La construction du système CK-Cartography s'appuie sur un méta-modèle présentant les concepts et leurs relations. Ce méta-modèle est basé sur une approche ontologique permettant de définir rigoureusement les concepts mis en jeu. Dans ce cadre, les ontologies noyaux COOK (*Core Ontology of Know-How and Knowing-That*) (Ghrab et al., 2016) et COOP (*Core Ontology of Organization's Processes*) (Turki et al., 2016) sont utilisées. Leur intérêt est la définition des concepts les plus généraux pour la cartographie des savoirs et des processus. Cette ontologie est indépendante du domaine traité. Le méta-modèle que nous proposons est représenté dans la Figure 1 par un diagramme de classes UML.

Le méta-modèle de cartographie des savoirs permet de représenter:

- les processus de l'organisation,
- les actions (individuelles ou collectives) mobilisées dans ces processus,

- les savoir-faire et les savoirs factuels identifiés dans ces processus et qui sont nécessaires pour la réalisation des actions,
- les acteurs et les collectifs intervenant dans les processus de l'organisation,
- les objectifs organisationnels de l'organisation.

Le méta-modèle propose un ensemble minimal de concepts requis et définis pour spécifier la cartographie des savoirs. Ces concepts pourraient être étendus pour répondre à des besoins spécifiques. Ce méta-modèle fait ressortir, principalement, la différence entre un savoir-faire et un savoir factuel par la proposition de concepts permettant de définir la nature de chaque concept (un savoir-faire est une disposition et un savoir factuel est un état de croyance). Un savoir-faire peut être composé de savoir-faire et de savoirs factuels qui peuvent être individuels ou collectifs. Les savoir-faire permettent de réaliser des actions individuelles et collectives. Un processus de l'organisation est une action de l'organisation qui a pour agent une organisation composée de plusieurs personnes. Une organisation est un type particulier de collectif. Selon Kassel et al. (2012), un collectif est un groupe d'humains unifiés par une intention jointe celle de former un groupe capable d'agir. Des objectifs organisationnels sont définis et validés pour l'organisation. Afin d'atteindre ces objectifs, les acteurs de l'organisation réalisent des actions en utilisant les savoir-faire et les savoirs factuels de l'organisation mobilisés dans les processus. Les processus simples sont distingués des processus composés. Un processus simple ne peut pas être composé d'autres processus contrairement au processus composé.

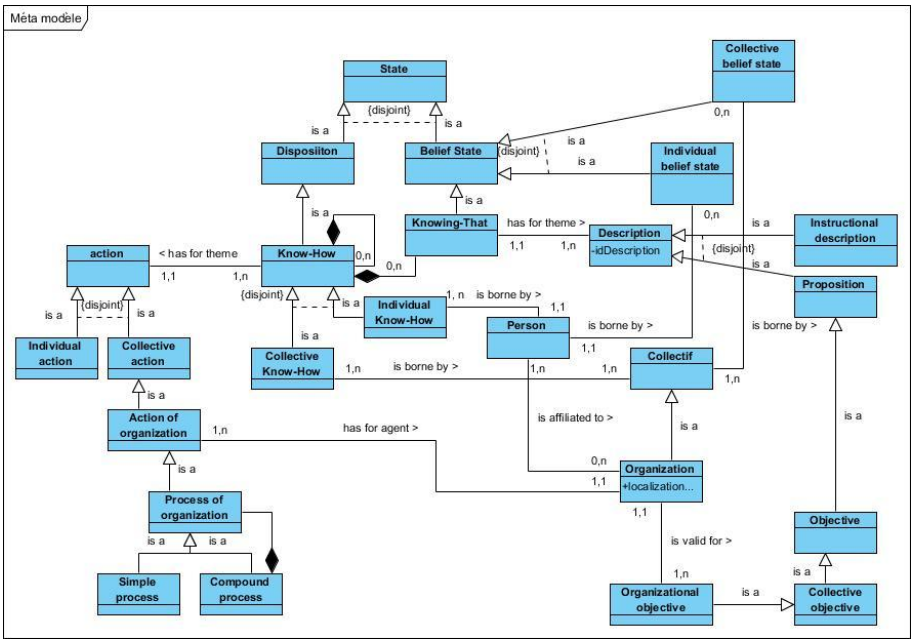


Figure 1 : Méta-modèle de la cartographie des savoirs

4. Langage graphique iconique

Pour la proposition de ce langage, nous nous basons principalement sur les techniques de visualisation d'information, la théorie de Gestalt, la psychologie cognitive et la sémiologie graphique (Bertin, 1999). Ce langage est responsable de la représentation visuelle des concepts du méta-modèle. Chaque concept est associé à un nœud qui peut être sous forme d'icône ou de pictogramme ou de forme géométrique. Il est possible d'avoir un nœud composé de ces trois structures afin de mettre en valeur certaines caractéristiques d'un concept (interne ou externe, collectif ou individuel, partagé ou non partagé). Le choix de ces structures visuelles est assuré par les techniques de visualisation d'information, la théorie de Gestalt, la psychologie cognitive et la sémiologie graphique. Les structures visuelles servent de base à une représentation commune partagée entre les différents acteurs de l'organisation sous la devise de Neurath: « les mots divisent, les images unissent».

Une forme peut être un cercle, un ovale, un rectangle ou un rectangle aux coins arrondis. Un cercle fait référence à un savoir-faire et un ovale fait référence à un savoir factuel. Un rectangle fait référence à un processus de l'organisation et un rectangle aux coins arrondis fait référence à une action (qui peut être individuelle ou collective). Comme nous avons indiqué dans le méta-modèle, un processus de niveau n est composé d'autres processus de niveau n+1. Une couleur est définie en fonction du niveau de granularité du processus (autant de couleurs définies que de niveau de granularité). Le niveau de granularité des processus dépend du terrain d'application étudié.

Un pictogramme utilisé dans la cartographie peut être une icône, un pictogramme en exposant ou un pictogramme en indice (Tableau 1). Les pictogrammes en exposant ou en indice sont utilisés avec d'autres formes et icônes afin de caractériser le référent. Le pictogramme en exposant réfère à un caractère individuel ou collectif d'un concept et le pictogramme en indice réfère à un caractère interne ou externe.

	Icône					**Pictogramme en exposant**		**Pictogramme en indice**	
Pictogramme									
Désignation	*Organisation*	*Acteur*	*Collectif*	*Support papier*	*Support numérique*	*Individuel*	*Collectif*	*Interne*	*Externe*

Tableau 1 : L'ensemble des pictogrammes utilisés dans la cartographie des savoirs et leur signification

Un savoir-faire peut être interne ou externe à l'organisation. Un savoir-faire interne est détenu par un membre affilié à l'organisation et un savoir-faire externe est détenu par un membre extérieur à l'organisation. Le savoir-faire est utilisé dans l'organisation pour la réalisation de certaines actions.

Pour empêcher la volatilité de ces savoir-faire et identifier le type de chaque savoir-faire et sa localisation, la distinction entre la dimension interne/externe est intéressante. Ces deux types de savoir-faire sont représentés par un composant graphique complexe composé du cercle représentant le savoir-faire et du pictogramme en indice comme mentionné dans le tableau1.

De même, un savoir-faire peut être tacite détenu par une personne ou un collectif ou explicité sur un support papier ou numérique. Dans un contexte organisationnel, un savoir-faire peut être partagé ou non partagé. En particulier, un savoir-faire tacite est partagé via des discussions informelles entre les membres de l'organisation mais n'est pas encore explicité sur un support physique. Un jeu de couleurs est proposé afin de mettre en œuvre ces types de savoir-faire. La couleur rose claire réfère à un savoir-faire tacite non partagé. La couleur jaune claire réfère à un savoir-faire tacite partagé. La couleur bleue claire réfère à un savoir-faire explicité non partagé et la couleur verte claire réfère à un savoir-faire explicité partagé.

5. Etude de cas

Le contexte applicatif du prototype CK-Cartography est l'Association de Sauvegarde des Handicapés Moteurs de Sfax (ASHMS) dont nous donnons une brève description de son fonctionnement. Nous détaillerons ensuite l'expérimentation de CK-Cartography dans l'ASHMS.

5.1. Présentation de l'ASHMS

L'ASHMS est un organisme à but non lucratif dont la mission principale est de venir en aide aux personnes handicapées et démunies de la Tunisie. L'ASHMS est administrée exclusivement par des bénévoles et bénéficie des subventions. Ses activités d'aide sont financées par des dons de la communauté et des contributions de ses partenaires. Les professionnels de santé rendant des services aux enfants handicapés sont des bénévoles et sont affiliés souvent à l'hôpital universitaire Habib Bourguiba de Sfax ou à la faculté de Médecine. Parmi ces professionnels de santé, il existe certains qui ont leur propre cabinet. Dans l'ASHMS, différents processus sont identifiés. Dans ce papier, nous nous intéressons particulièrement au processus de prise en charge précoce des enfants IMC (Infirmité Motrice Cérébrale) qui consiste à évaluer l'état de santé de l'enfant tous les 3 mois lors d'une réunion de staff comportant les différents médecins et techniciens de santé qui ont contribué à l'examen et à l'évaluation de cet enfant durant ces 3 mois. Afin de prendre une décision collective sur l'arrêt ou la poursuite de la rééducation de l'enfant IMC, plusieurs savoir-faire, savoirs factuels et informations doivent être échangés entre les différents professionnels de santé pour avoir une idée globale sur le développement de l'état de l'enfant sur tous les plans (neuropédiatrique, néonatalogique, kinésithérapique, ergothérapique, orthophonique, pédopsychiatrique, etc.). Lors de cette réunion et même avant son déroulement, les professionnels de santé peuvent rencontrer plusieurs difficultés qui peuvent empêcher la qualité de soin et d'évaluation de l'enfant IMC et la prise de décision collective. En fait, plusieurs informations peuvent être manquantes dans les dossiers médicaux et les comptes rendus. Certains professionnels peuvent ne pas participer à la réunion du staff vu des engagements qui les empêchent (engagement professionnel dans leur organisation d'origine, changement de poste etc.) surtout que la plupart de ses professionnels sont des bénévoles. Même si chaque professionnel de santé a des connaissances se rapportant à sa spécialité, lors de la réunion de staff, il doit avoir aussi d'autres informations et connaissances se rapportant aux différentes autres spécialités afin de mieux juger l'état de l'enfant IMC.

5.2. Expérimentation de CK-Cartography dans l'ASHMS

Nous distinguons deux types de professionnels de santé dans l'ASHMS : le détenteur du savoir et l'utilisateur du savoir. Le détenteur du savoir est un médecin ou technicien de santé. C'est lui qui crée le savoir et responsable de sa gestion. L'utilisateur du savoir est tout professionnel de santé consultant le savoir soit pour la compréhension des savoirs globaux dans une spécialité donnée, soit pour l'utilisation de ce savoir lors de la réunion du staff médical , soit par un apprenant pour bénéficier de l'expérience de leurs maîtres (apprentissage maître-apprenti).

La cartographie des savoirs générée par CK-Cartography donne une vue globale et détaillée des différents processus de l'organisation. Dans l'ASHMS, nous distinguons entre 4 niveaux de granularité des processus (Turki et al., 2012): les processus FLP (*First Level Process*), les processus ILP (*Intermediate Level Process*)

et les processus OP (*Organizational Process*). Les processus FLP ne font pas partie d'autres processus. Ce sont des processus simples. Les processus ILP sont des processus qui font partie d'un autre processus. Un ILP peut se composer d'un nombre défini de processus qui dépend de la complexité des processus du niveau supérieur. Nous distinguons les processus TLP (*Third Level Process*) des processus SLP (*Second Level Process*). Le processus OP représente des actions délibérées de l'organisation ayant pour agent un individu ou une unité de l'organisation. En fonction du niveau de granularité des processus de l'ASHMS, quatre couleurs sont choisies pour référer chaque processus de granularité i. La couleur jaune foncée est utilisée pour référer à un processus de type FLP. La couleur verte foncée est utilisée pour référer à un processus de type SLP. La couleur bleue foncée est utilisée pour référer à un processus de type TLP et la couleur rose foncée est utilisée pour référer à un processus de type FLP.

La relation qui existe entre deux processus de types différents est une relation de type "contribute to". Nous étudions dans la figure 2 le "processus de prise en charge médical d'un enfant IMC ayant la forme hémiplégie". Ce processus est composé de processus OP (par exemple, "processus de prise en charge d'un enfant IMC en kinésithérapie (IMC de type hémiplégie)", "processus de prise en charge d'un enfant IMC en ergothérapie (IMC de type hémiplégie)", "processus de prise en charge d'un enfant IMC en neuropédiatrie (IMC de type hémiplégie)", etc.).

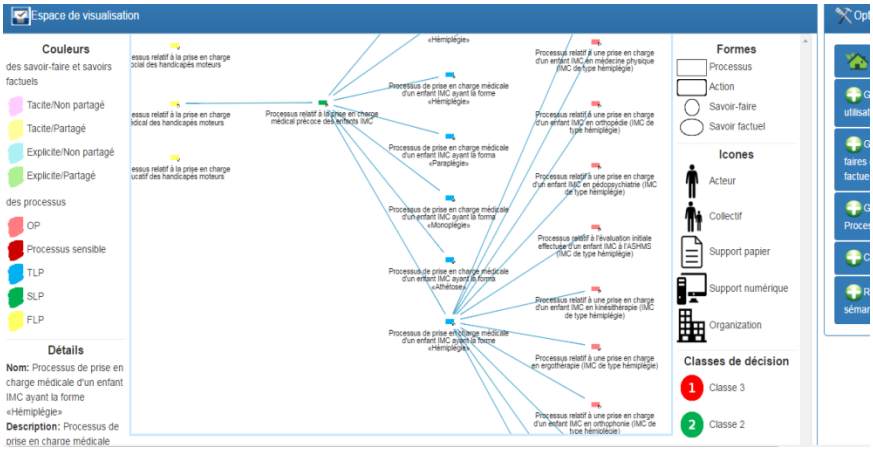


Figure 2. Capture d'écran du passage de la cartographie des processus d'un niveau à un autre niveau de granularité plus fine (FLP-SLP-TLP-OP)

Le clic sur le "processus relatif à une prise en charge d'un enfant IMC en neuro-pédiatrie (IMC de type Hémiplégie)" de la carte de processus (Figure 2) permet d'afficher la liste des savoir-faire et des savoirs factuels mobilisés dans ce processus. Parmi les savoir-faire identifiés, nous citons le "savoir détecter l'hypertonie et l'hypotonie" qui est un savoir-faire composé (ayant une taille plus grande que les

autres savoir-faire) d'autres savoir-faire et savoirs factuels. Ce savoir-faire est un savoir-faire tacite partagé par le staff médical.

Dans la figure 3, "savoir évaluer la motricité spontanée" est un savoir-faire composé des savoirs factuels ("savoir le développement de la capacité d'acquisition psychomotrice", "savoir l'anomalie neurologique pour le développement moteur") et des savoir-faire ("savoir chercher un mouvement anormal", "savoir évaluer le tonus axial" et "savoir détecter l'hypertonie et l'hypotonie").

Le savoir-faire "évaluer la motricité spontanée" est un savoir-faire tacite non partagé. Il est souvent partagé par compagnonnage (du maître à l'apprenti). Ce savoir-faire a pour thème l'action "évaluer la motricité spontanée" qui est une action individuelle interne. Ce savoir-faire est détenu par le médecin neurologue qui est affilié à l'hôpital universitaire et est capable de réaliser plusieurs actions internes à l'ASHMS (évaluer l'état d'éveil, évaluer le tonus axial, etc.).

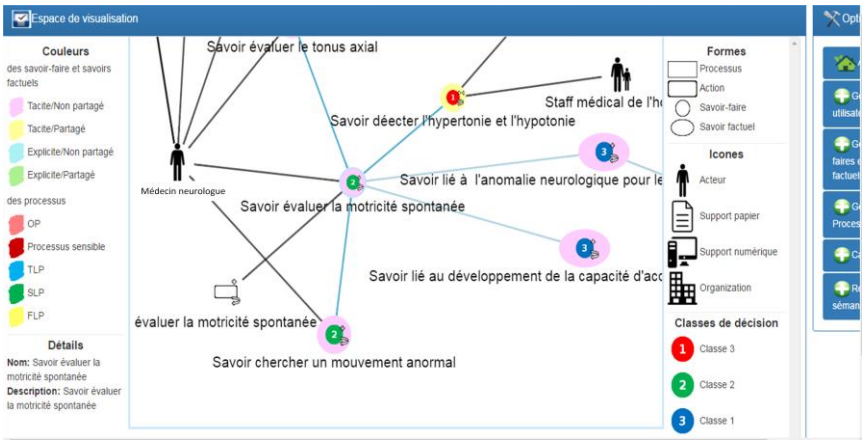


Figure 3. Capture d'écran de la carte de savoir-faire et de savoirs factuels relative au "savoir évaluer la motricité spontanée"

En cas d'absence d'un professionnel de santé lors de la réunion du staff, le professionnel peut via cette carte échanger avec les autres membres du staff médical son compte rendu sur l'évaluation de l'enfant IMC et participer à distance à la prise de décision collective (arrêt ou poursuite de la rééducation de l'enfant IMC). Par conséquent, lors de la réunion du staff, le staff médical visualise aisément les professionnels de santé qui ont participé à la rééducation de l'enfant IMC (Figure 4). Nous distinguons donc le professionnel de santé détenteur du savoir (du ou des) professionnel(s) de santé utilisateur de ce savoir.

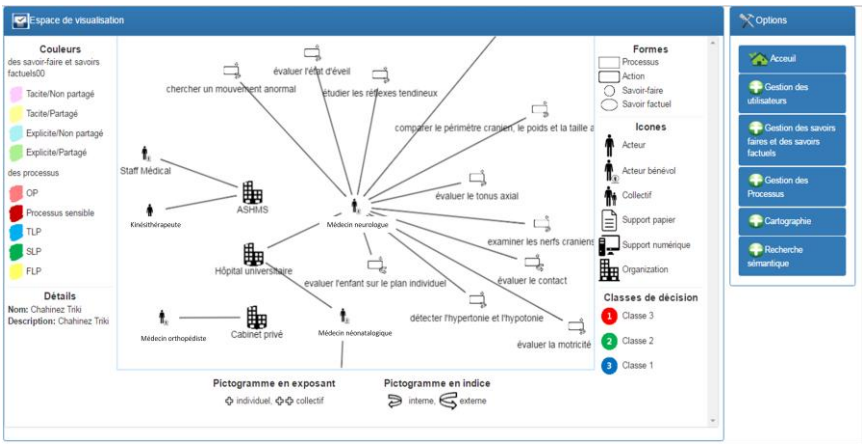


Figure 4 : Carte de personnel

6. Démarche pour la construction de la cartographie des savoirs

Dans cette section, nous présentons la démarche proposée pour construire et valider la cartographie générée par CK-Cartography ainsi que les enseignements tirés. La construction de cette méthode est le résultat de plusieurs expérimentations menées sur des processus de prise en charge des enfants IMC. Ces expérimentations ont été exploitées pour consolider la méthode.

6.1. Démarche d'élaboration de CK-Cartography

La cartographie des savoirs produite par CK-Cartography est générée selon une démarche de cartographie des savoirs composée de quatre étapes : (i) l'analyse conceptuelle des savoir-faire et des savoirs factuels, (ii) la proposition d'une ontologie noyau COOK, (iii) l'extension de la méthode d'identification des connaissances cruciales (Saad et Ghrab, 2016) et (iv) la proposition d'un langage graphique iconique (Ghrab, 2016 ; Ghrab et al., 2017).

La première étape de la démarche consiste à analyser conceptuellement les notions de savoir-faire et de savoir factuel. Les savoir-faire et les savoirs factuels sont deux types de connaissance. Le savoir-faire est lié à la notion d'action tandis que le savoir factuel est lié à la notion de description. Un savoir-faire est la disposition à réaliser une action et un savoir factuel est un état de croyance relatif à une description qui peut être factuelle ou prescriptive.

La proposition de l'ontologie noyau COOK est basée sur l'étape précédente de l'analyse conceptuelle des savoir-faire et des savoirs factuels. L'objectif de COOK (Ghrab et al., 2016) est de proposer une ontologie noyau indépendante du domaine permettant des définitions rigoureuses des savoir-faire et des savoirs factuels.

La troisième étape de la démarche permet d'identifier les savoir-faire et les savoirs factuels à visualiser dans la cartographie des savoirs ainsi que les éléments qui sont en rapport avec les savoir-faire et les savoirs factuels (acteur, support, organisation, type de savoir, processus, type de processus, etc.) Cette méthode encapsule la méthode d'identification des processus sensibles de l'organisation (Ghrab et Saad, 2016) où sont mobilisés les savoirs cruciaux. Nous nous basons principalement sur une approche processus de cartographie des connaissances.

Et finalement, la dernière étape est la proposition d'un langage graphique iconique permettant aux utilisateurs de CK-Cartography de créer un langage visuel commun leur permettant de mieux transférer et interpréter aisément et facilement les savoirs.

6.2. Enseignement tirés

L'approche de construction de CK-Cartography que nous proposons est une approche guidée par les intéressés représentés par les utilisateurs futurs du système CK-Cartography (les professionnels de santé). Ces professionnels sont affiliés à des organisations différentes (ASHMS, hôpital universitaire, faculté de médecine, autre association, cabinet privé, etc.). L'implication de ces utilisateurs n'est pas assurée uniquement lors du développement du système mais aussi dès les premières phases de conception et de modélisation du système afin de garantir leur satisfaction. Des entretiens avec les différents professionnels de santé ont été menés au cours et après le développement de CK-Cartography afin d'étudier leur degré de satisfaction. D'autres évaluations ergonomiques ont été menées. Leur but est de déterminer l'utilité (adéquation du système avec les besoins de l'utilisateur) et l'utilisabilité (simplicité d'utilisation, ergonomie des interfaces, clarté des fonctionnalités du système, guidance de l'utilisateur lors de l'utilisation du système) de CK-Cartography (Figure 5), proposer des recommandations, des spécifications et des maquettes permettant une conception du prototype plus adaptée aux utilisateurs. Ces évaluations sont basées sur une évaluation heuristique et des tests utilisateurs qui sont des méthodes en ergonomie des IHM (Interface Homme Machine).

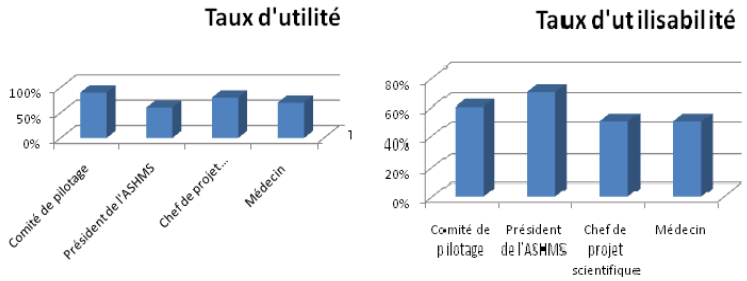


Figure 5 : Degré de satisfaction des professionnels de santé de CK-Cartography

L'usage de CK-Cartography permet de gagner du temps dépensé à chercher les informations nécessaires au suivi et à l'évaluation des enfants IMC lors de la réunion du staff médical. Ces informations peuvent être stockées dans des supports différents (compte rendu du médecin, des radios et des analyses effectués à l'extérieur de l'ASHMS, des dossiers médicaux).

A cause des informations manquantes dans les comptes rendus de chaque examen ou lors de la réunion du staff, les professionnels de santé ont du mal à évaluer l'état des enfants IMC et trouvent des difficultés à interpréter et échanger ces informations. Grâce à CK-Cartography, ce problème est résolu.

7. Conclusion

Dans ce papier, nous avons proposé le système d'informations et de connaissances CK-Cartography de transfert des savoirs factuels et des savoir-faire. Ce système est basé sur une démarche pour la construction de la cartographie des savoirs composé de quatre étapes : analyse conceptuelle des savoir-faire et des savoirs factuels, proposition d'une ontologie noyau COOK, extension de la méthode d'identification des connaissances cruciales de Saad et al. (2009) et proposition d'un langage graphique iconique qui a été détaillé.Ce langage est basé sur un jeu de forme, de couleur et de pictogrammes combinés afin de donner une visualisation plus intuitive. Un méta-modèle de cartographie des savoirs est proposé. Il est basé sur l'ontologie COOK pour la définition rigoureuse des concepts ainsi que les relations définies entre ces concepts. CK-Cartography est défini comme un outil pour l'unification des communications entre les individus de l'organisation.

Expérimenté dans l'ASHMS, CK-Cartography est utilisé comme un outil d'aide à la visualisation des savoir-faire et des savoirs factuels pour les professionnels de santé dans le but d'améliorer la qualité de prise en charge des enfants IMC.

Comme travaux futurs, nous souhaiterons intégrer d'autres modules assurant l'apprentissage maître-apprenti et la recherche sémantique sous forme visuelle. D'autres concepts devraient être aussi pris en considération pour enrichir notre méta-modèle de cartographie des savoirs (compétence, savoir-faire inter-organisationnel, savoir factuel inter-organisationnel, savoir déclaratif, savoir procédural, savoir factuel partagé et non partagé, etc.) et d'autres concepts devraient être mieux approfondis (savoir-faire partagé, savoir-faire non partagé).

Bibliographie

Arduin P.-E.,Grundstein M., Rosenthal-Sabroux C. (2015). *Système d'information et de connaissance*. Edition ISTE.

Balaid A. S. S., Zibarzani M., Rozan M. Z. A. (2013). A comprehensive review of knowledge mapping techniques. *Journal of Information systems research and innovation*, vol 3, p. 71–76.

Bertin J. (1999). *Sémiologie graphique : les diagrammes, les réseaux, les cartes*. Paris, Ecole des Hautes Etudes en Sciences Edition.

Bresciani S., Eppler M. (2013). Knowledge Visualization for Social Entrepreneurs. In IEEE Proceedings of the 16th *International Conference Information Visualization* IV13. p. 319–324.

Crampes M., Ranwez S., Villerd J. (2008). Cartographie sémantique auto-organisée d'un référentiel de connaissances partagé. 19èmes Journées Francophones d'*Ingénierie des Connaissances* (2008). Nancy, France, p. 161–172.

Dorze A. L., Garcia L., Genest D., Loiseau S. (2014). Synthesis of cognitive maps and applications. 26th *IEEE International Conference on Tools with Artificial Intelligence* (ICTAI). Limassol, Cyprus. p. 291–298.

Gandon F. (2008). *Graphe RDF et leur manipulation pour la gestion des connaissances*. Mémoire d'habilitation à diriger les recherches, Université de Nice.

Ghrab S., Saad I., Gargouri F., Kassel G. (2017). A decision support system CK-Cartography for knowledge cartography. 3rd *International Conference on Knowledge Management, Information and Knowledge Systems* (KMIKS 2017), Hammamet, Tunisia, 20-22 April. p. 91-102.

Ghrab S. (2016). *Elaboration d'une cartographie multicritère pour l'aide à la caractérisation et l'évaluation des connaissances médicales*. Thèse de doctorat en Informatique, Université de Sfax-Université de Picardie Jules Verne.

Ghrab S., Saad I. (2016). Identifying crucial Know-How and Knowing-That for medical decision support. *International Journal of Decision Support System Technology*, vol. 8, n° 4, p. 14-33.

Ghrab S., Saad I., Kassel G., Gargouri F. (2016). A Core Ontology of Know-How and Knowing-That for improving knowledge sharing and decision making in the digital age. *Journal of Decision Systems*, vol. 10, n° 10, p. 1-14.

Hao J., Yan Y., Gong L., Wang G., Lin J. (2014). Knowledge map-based method for domain knowledge browsing. *Decision Support Systems*, vol. 61, p. 106-114.

Kassel G., Turki M., Saad I., Gargouri F. (2012). From collective actions to actions of organizations: an ontological analysis. In *Symposium Understanding and Modelling Collective Phenomena* (UMoCoP 2012), Birmingham, England.

Saad I., Grundstein M., Rosenthal-Sabroux C. (2009). Une méthode d'aide à l'identification des connaissances cruciales pour l'entreprise. *Revue Systèmes d'Information et Management* (SIM), vol. 14, n° 3, p. 43–78.

Saad I., Rosenthal-Sabroux C., Gargouri F. (2017). Knowledge sharing and decision making in the age of digital. *Journal of Decision Systems*, vol. 26, n° 2.

Sellin C. (2011). *Des organisations centrées processus aux organisations centrées connaissance : la cartographie de connaissances comme levier de transformation des organisations. Le cas de la démarche de « transfert de savoir-faire »*. Thèse de doctorat en sciences de gestion, Ecole centrale de Paris.

Tricot C.(2006). *Cartographie sémantique : des connaissances à la carte*. Thèse de doctorat en informatique, Université Haute Savoie.

Turki M., Saad I., Gargouri F., Kassel G. (2012). A Decision Support System for Identifying Sensitive Organization's Processes. *Journal of decision Systems*, vol. 21, n° 4, p. 275-290.

Turki M., Kassel G., Saad I., Gargouri F. (2016). A Core ontology of business processes based on DOLCE. *Journal on Data Semantics*, vol. 5, n° 3, p. 165-177.

Wickel M., Schenkl S., Schmidt D., Hense J., Mandl H., Maurer M. (2013). Knowledge structure maps based on multiple domain matrices. In Impact: The *Journal of Innovation Impact*, vol. 5, p. 5-16.

Modèles : concepts et ingénierie

Approche guidée pour l'anonymisation de bases de données

Feten BenFredj[1], Nadira Lammari[1], Isabelle Comyn-Wattiau[2]

1. CEDRIC-CNAM, 2 Rue Conté, 75003 Paris, France
2. ESSEC Business School, 1 Av. Bernard Hirsch, 95021 Cergy, France
fetenbf@yahoo.fr, ilham-nadira.lammari@cnam.fr, wattiau@essec.edu

RESUME. L'anonymisation des données personnelles requiert l'utilisation d'algorithmes complexes permettant de minimiser le risque de ré-identification tout en préservant l'utilité des données. Dans cet article, nous décrivons une approche fondée sur les modèles qui guide le propriétaire des données dans son processus d'anonymisation. Le guidage peut être informatif ou suggestif. Il permet de choisir l'algorithme le plus pertinent en fonction des caractéristiques des données mais aussi de l'usage ultérieur des données anonymisées. Le guidage a aussi pour but de définir les bons paramètres à appliquer à l'algorithme retenu. Dans cet article, nous nous focalisons sur les algorithmes de généralisation de micro-données. Les connaissances liées à l'anonymisation tant théoriques qu'expérimentales sont stockées dans une ontologie.

ABSTRACT. Personal data anonymization requires complex algorithms aiming at avoiding disclosure risk without losing data utility. In this paper, we describe a model-driven approach guiding the data owner during the anonymization process. The guidance may be informative or suggestive. It helps the data owner in choosing the most relevant algorithm given the data characteristics and the future usage of anonymized data. The guidance process also helps in defining the best input values for the algorithms. In this paper, we focus on generalization algorithms for micro-data. The knowledge about anonymization is composed of both theoretical aspects and experimental results. It is managed thanks to an ontology.

MOTS-CLES : guidage, sécurité, ontologie, méthodologie, respect de la vie privée, anonymisation, approche guidée par les modèles.

KEYWORDS: guidance, security, ontology, methodology, privacy, anonymization, model-driven approach.

1. Introduction

Le partage des données au-delà des frontières même de l'organisation s'est accentué, par exemple, par l'engagement des pays sur la voie de l'ouverture des données publiques, plus connue sous le nom d' «open data». Cette situation soulève la question du risque de divulgation de données sensibles, et plus particulièrement, le risque de violation de la vie privée via l'utilisation de données personnelles. La

norme ISO/TS 25237:2008 définit l'anonymisation comme «un processus qui supprime l'association entre l'ensemble de données identifiant et le sujet des données». C'est un processus complexe, notamment parce qu'il tente de satisfaire deux objectifs contradictoires que sont : l'utilité des données (c'est-à-dire leur qualité) et leur sécurité (c'est-à-dire leur confidentialité). Par conséquent, les éditeurs de données sont toujours à la recherche d'une solution qui réponde au mieux à la confidentialité et à l'utilité de leurs données. Leur solution émerge après des prises de décision à différentes phases du déroulement de leur tâche. En effet, ils sont amenés entre autres à sélectionner un algorithme d'anonymisation, à opter pour un paramétrage adéquat de cet algorithme et à juger de la qualité du rendu après application du procédé ainsi paramétré. Ils sont donc engagés dans un processus de décision qui s'appuie sur leur connaissance du domaine.

Les outils existants, de par leur opacité et leur manque de guidage dans le choix et le paramétrage des algorithmes, ne simplifient pas cette activité pour un professionnel ayant une faible expertise dans le domaine. D'un point de vue académique, nous avons aussi constaté l'absence d'approches guidées pour l'anonymisation bien que la littérature abonde d'articles de recherche sur les algorithmes d'anonymisation.

Ces constats ont motivé notre démarche de création d'une ontologie de domaine pour l'anonymisation de micro-données[1] ainsi que d'une approche guidée s'appuyant sur cette ontologie. L'ontologie produite (BenFredj *et al.,* 2015), que nous avons nommée OPAM, permet de capitaliser les connaissances du domaine. Cependant, elle ne stocke qu'une portion d'expertise du domaine. En effet OPAM, n'a été, pour l'instant, instanciée que par les connaissances récoltées sur la technique de généralisation de micro-données. Par conséquent, l'approche, que nous proposons dans cet article et que nous nommons MAGGO (Méthodologie pour une Anonymisation par Généralisation Guidée par une ontologie), sert de guide pour un professionnel dans sa prise de décision lors d'une anonymisation par généralisation de micro-données. Cela n'enlève rien à la généricité de notre approche. En effet, elle peut être instanciée pour une autre technique.

Après un rapide état de l'art (section 2), nous décrivons l'approche générale (section 3), puis ses étapes détaillées (section 4). En section 5, nous illustrons l'approche sur un exemple. Enfin, nous concluons en section 6 et présentons quelques axes de recherche future.

2 Etat de l'art

Plusieurs techniques d'anonymisation existent avec des degrés de fiabilité et des contextes d'applicabilité variables. Ces contextes sont caractérisés, entre autres, par l'usage souhaité des données (par exemple pour tester un logiciel ou encore pour publier les données à des fins d'analyse) et par le type de données à anonymiser (micro ou macro données tabulaires, données spatio-temporelles, graphes, images,

[1] Données atomiques décrivant des individus (Hand, 1992)

textes, etc.). Le degré de fiabilité est en lien direct avec le risque de ré-identification des données anonymes. Cependant, face à l'évolution des technologies de l'information qui rendent possible le lien entre des données de différentes sources, il est quasiment impossible d'effectuer une anonymisation qui garantirait un risque de ré-identification nul.

Les techniques d'anonymisation peuvent être classées en deux catégories : les techniques perturbatrices et les techniques non perturbatrices (Patel et Gupta, 2013). La première catégorie représente les procédures dans lesquelles les données résultantes ne sont pas dénaturées, c'est-à-dire que les données sont vraies mais qu'elles peuvent manquer de détails, alors que les données de la deuxième catégorie sont dénaturées, c'est-à-dire inexactes, ce qui n'empêche pas leur usage à des fins de test ou de statistique par exemple. La technique de suppression consiste à retirer des données de la table pour éviter leur divulgation. C'est une technique non perturbatrice. La technique de recodage global (« global recoding ») s'applique à toutes les valeurs d'un attribut afin d'uniformiser au plus les enregistrements et donc de diminuer le risque de ré-identification. Ainsi, on peut remplacer l'âge d'un individu par un intervalle. La technique de généralisation consiste à remplacer des valeurs par des valeurs plus générales (Samarati, 2001) : les données sont vraies, mais moins précises. La généralisation est appliquée à un ensemble d'attributs formant un quasi-identifiant (QI). Elle nécessite la définition d'une hiérarchie pour chaque attribut composant le QI. L'âge peut être généralisé à l'aide d'intervalles de valeurs de plus en plus grands vers la racine de la hiérarchie. Généraliser consiste à remplacer une valeur par son ancêtre direct dans la hiérarchie de généralisation, à chaque étape de la généralisation. Ainsi, on peut appliquer une seule étape de généralisation à l'attribut Ville et deux étapes de généralisation à l'attribut Age. Le « data swapping » (Fienberg et McIntyre, 2004), consiste à permuter les valeurs d'un même attribut entre des paires d'enregistrements. La micro-agrégation (Defays et Nanopoulos, 1993) répartit les données originales en groupes homogènes. Par la suite, les valeurs originales sont remplacées par la moyenne ou la médiane du groupe auquel elles appartiennent. La technique de bruit aléatoire (« random noise ») (Brand, 2002) s'applique à un seul attribut à la fois. Elle fonctionne en ajoutant ou en multipliant chaque valeur de l'attribut à anonymiser par une variable aléatoire. Chacune de ces techniques a donné lieu à un ou plusieurs algorithmes. Par exemple, la généralisation peut s'appliquer de différentes façons et des dizaines d'algorithmes ont été proposés dans cette catégorie.

Ainsi, il existe une grande variété de techniques d'anonymisation et encore plus d'algorithmes qui les mettent en œuvre. Des comparaisons de techniques sont proposées (Ilavarasi *et al.*, 2013, Fung *et al.*, 2010). Certaines sont certes orientées usage mais restent non accessibles à des éditeurs de données avec de faibles compétence dans le domaine. De plus, les algorithmes associés aux techniques ne sont accessibles qu'à travers les publications de recherche. Leur spécification se rapproche du code de programmation. Ils sont, le plus souvent, partiellement illustrés à l'aide d'exemples. Leurs principes fondamentaux sont décrits textuellement. Par conséquent, ils ne sont compréhensibles que par des informaticiens ou des professionnels ayant des compétences en programmation.

Il existe aussi des logiciels d'anonymisation[2] (Poulis *et al*., 2014 ; Xiao *et al*., 2009 ; Dai *et al.*, 2009). Le plus souvent, ils sont opaques. Même s'ils proposent plusieurs techniques, ils mettent en œuvre, en général, un seul algorithme par technique sans mentionner lequel. La plupart de ces outils ne fournissent pas de guidage dans le choix de la technique et de l'algorithme. Ils n'offrent pas d'aide au paramétrage des algorithmes proposés. Le guidage est réduit à l'application de métriques sur les données anonymisées qui permettent à l'éditeur de données d'évaluer notamment le risque résiduel et la dégradation due à l'anonymisation.

L'état de l'art comprend aussi de nombreuses métriques permettant d'évaluer la qualité des données anonymisées, en termes de perte d'information ou de précision, ou le risque de ré-identification (Fung *et al*., 2010).

Enfin, à notre connaissance, à l'exception de notre ontologie OPAM (BenFredj *et al.*, 2015), il n'existe pas de base de connaissance dans laquelle le professionnel chargé de la désidentification des données pourrait rechercher les connaissances le guidant vers une anonymisation utile et préservant au mieux la vie privée. Il n'existe pas non plus de méthode qui puisse concrétiser le processus d'anonymisation de données tout en offrant des aides à la décision. Ainsi, dans cet article, nous définissons une approche d'aide à la décision permettant, à l'aide de l'ontologie OPAM, de guider l'éditeur de données dans le choix d'un algorithme et dans son paramétrage.

Dans la suite, nous présentons l'approche MAGGO en détaillant ses étapes principales.

3. Présentation générale de l'approche

L'anonymisation de données est une des mesures de sécurité qui peuvent être préconisées dans le cadre de la protection de la vie privée. Dès lors que cette mesure est décidée, le responsable de l'anonymisation doit concevoir et exécuter un processus de brouillage. Pour cela, il doit a) repérer les données identifiantes[3], quasi-identifiantes (QI)[4] et sensibles[5], b) proposer des techniques appropriées avec une orchestration adéquate. Il lui faut aussi, pour chaque technique, identifier

[2] PARAT est un exemple en ligne (http://www.privacyanalytics.ca/software/parat/).

[3] Un identifiant est un attribut ou un ensemble d'attributs qui désigne directement un individu (par exemple, un numéro de sécurité sociale, un prénom, un nom). Ce n'est pas nécessairement un identifiant au sens de la modélisation conceptuelle, puisqu'un prénom et/ou un nom peuvent être partagés par plusieurs individus. Toutefois, au sein d'un jeu de données, ce type d'information nominative peut facilement conduire à une ré-identification.

[4] Un quasi-identifiant (QI) est un ensemble d'attributs dont la sélectivité est telle qu'ils présentent un risque de ré-identification, Par exemple {sexe, code postal, date de naissance} forme un quasi-identifiant connu dans de nombreux ensembles de données. Ils sont suffisamment discriminants pour permettre de retrouver une seule personne dans une base de données

[5] Un attribut sensible représente les données que les individus ne veulent généralement pas publier, comme des informations médicales ou des salaires

l'algorithme à appliquer, trouver un paramétrage reflétant ses besoins et évaluer la qualité des données anonymisées en termes d'utilité et de sécurité en se conformant au cahier des charges de l'anonymisation. Ce processus comprend plusieurs points de décisions-clés dont la qualité affecte le résultat final. Il exige du responsable de l'anonymisation une grande maîtrise du domaine. Offrir une aide sur la totalité du processus exigerait des efforts considérables compte tenu de la variété des données susceptibles d'être brouillées (micro-données, données liées, données géographiques, etc.) et de la diversité des techniques existantes et des algorithmes de mise en œuvre de ces techniques. Dans cet article, nous nous focalisons sur une partie du processus d'anonymisation, c'est-à-dire une technique (la généralisation) et un type de donnée (les micro-données contenues dans une table relationnelle). En effet, nous proposons une approche guidée permettant, compte tenu d'un contexte d'anonymisation (défini dans un cahier des charges) de choisir l'algorithme de généralisation de micro-données le meilleur - au regard des exigences du cahier des charges - et de l'exécuter. Le meilleur algorithme est celui qui offre le meilleur compromis entre les exigences contradictoires de sécurité et d'utilité. Plus précisément, la recherche du compromis se fera en évaluant plusieurs algorithmes avec plusieurs combinaisons possibles de paramètres.

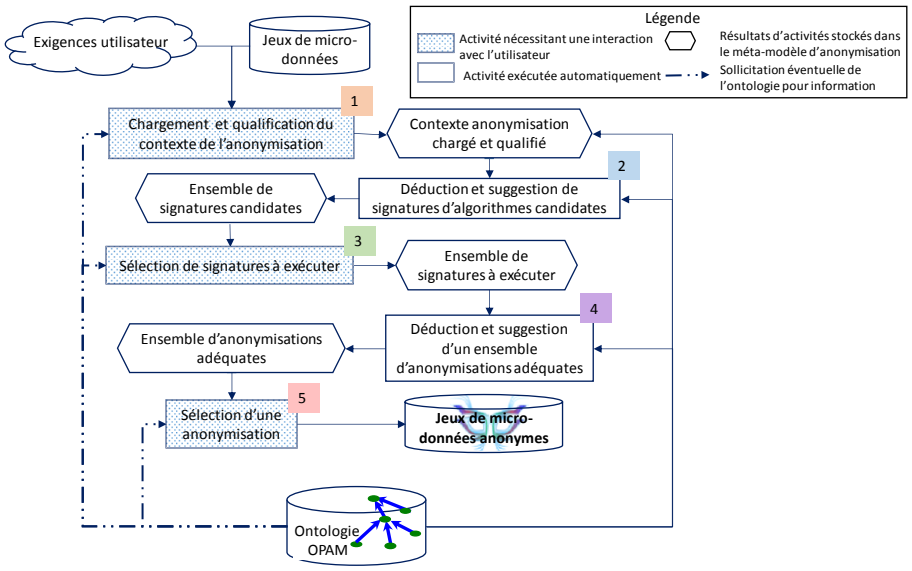


Figure 1. Les étapes de MAGGO

Pour aider l'utilisateur dans la spécification du contexte, dans la sélection de signatures et de solutions d'anonymisation, MAGGO met à disposition de l'utilisateur des connaissances nécessaires pour le rendre apte à décider. Ainsi, à chacune des étapes, MAGGO fait intervenir des connaissances expertes en vue d'un guidage suggestif ou informatif. Nous reprenons ces concepts de (Silver, 2006) : le

guidage suggestif guide l'utilisateur dans ses choix alors que le guidage informatif lui fournit des informations qui peuvent éclairer son choix. Dans notre cadre, le guidage suggestif aide l'éditeur de données dans la sélection de l'algorithme approprié tandis que le guidage informatif lui fournit des informations pour éclairer son choix sur un algorithme ou sur une technique.

Tableau 1. Type de guidage pour chaque étape

Etape	Activité	Guidage
1	Chargement et qualification du contexte de l'anonymisation	informatif
2	Déduction et suggestion d'un ensemble de signatures candidates	**suggestif**
3	Sélection de signatures à exécuter	informatif
4	Déduction et suggestion d'un ensemble d'anonymisations adéquates	**suggestif**
5	Sélection d'une anonymisation	informatif

Le tableau 1 récapitule les types de guidage offerts dans MAGGO selon l'étape. Ces connaissances sont rendues disponibles via OPAM. Le guidage de MAGGO est incrémental dans le sens où il est introduit à différents points de décisions clés tout au long du processus.

La notion de méta-modèle joue un rôle central dans notre approche. En effet, alors que l'ontologie met à disposition de l'approche les connaissances nécessaires à l'anonymisation, le méta-modèle (Fig. 2) réunit les abstractions conceptuelles des artefacts cibles et sources de notre approche.

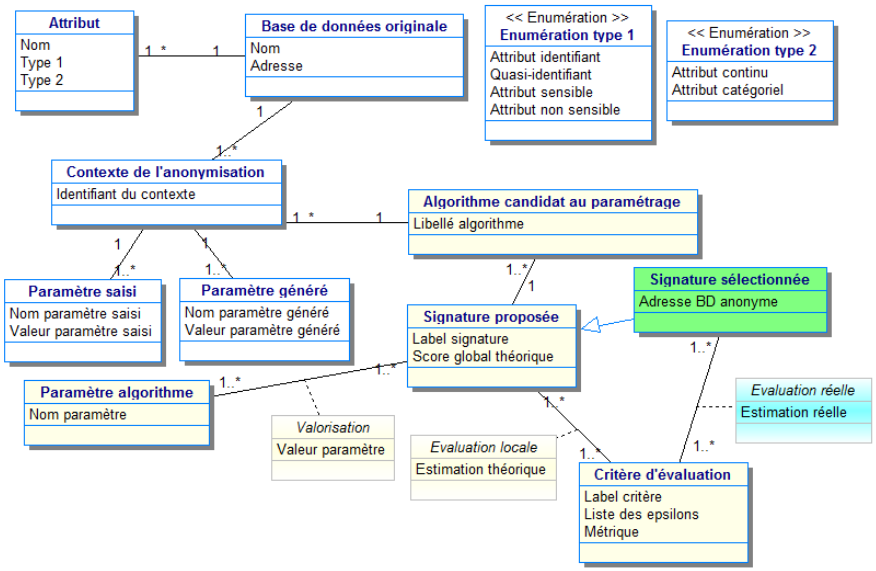


Figure 2. Le méta-modèle du processus d'anonymisation

La base de données originale est caractérisée par un ensemble d'attributs dont on définit le type 1 (identifiant, quasi-identifiant, sensible, non sensible) et le type 2

(continu ou catégoriel). Son contexte[6] est défini à partir de paramètres saisis (par l'utilisateur) ou générés (calculés automatiquement ou déduits des paramètres saisis). En fonction du contexte, on déduit des algorithmes candidats. Pour ces algorithmes, on déduit des signatures de paramètres à évaluer. L'évaluation peut être théorique, c'est-à-dire déduite des évaluations comparables contenues dans l'ontologie, ou réelle c'est-à-dire déduite d'une exécution de l'algorithme sur le jeu de données. Chaque couleur dans la figure correspond à l'étape de MAGGO au cours de laquelle les éléments correspondants sont sollicités.

Ainsi, l'exécution de la première étape de MAGGO permet d'instancier notre méta-modèle à l'aide des données relatives au cahier des charges. Cette instanciation correspond à une description du contexte de l'anonymisation ainsi qu'à sa qualification. Un enrichissement du modèle, par des données complémentaires issues de chacune des différentes étapes, est effectué.

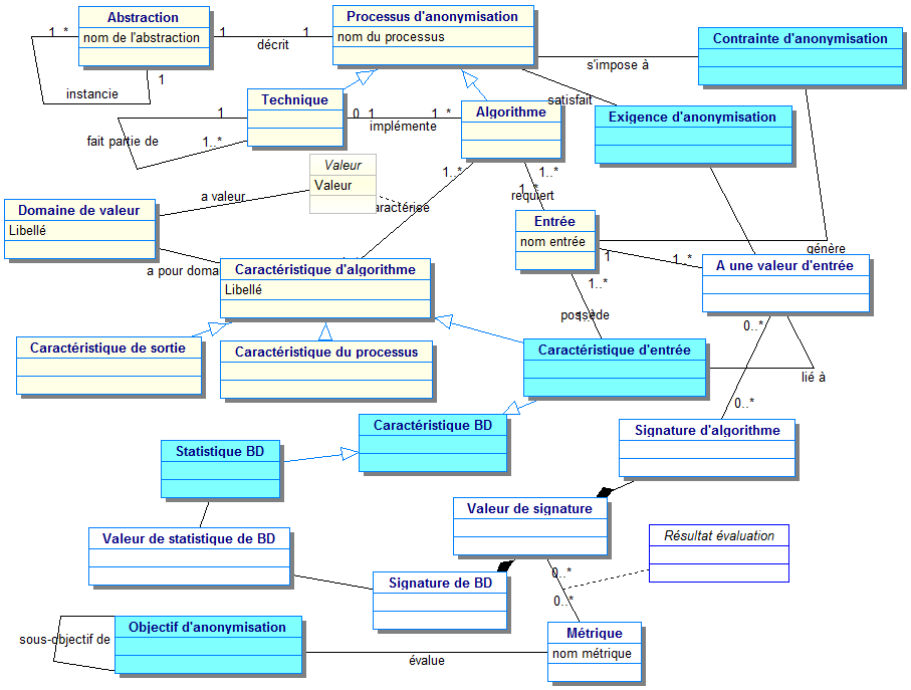


Figure 3. Méta-modèle sous-jacent de l'ontologie OPAM (extrait)

De plus, MAGGO exploite pour l'enrichissement du méta-modèle deux techniques statistiques : la technique d'aide à la décision multicritère Analytical Hierarchy Process (AHP) (Saaty et Sodenkamp, 2008) et la régression. La première est utilisée aux étapes 2 et 4 pour évaluer les résultats soumis à l'utilisateur. La

[6] Le tableau 2 fourni plus loin liste les paramètres de chaque catégorie. Par exemple, les attributs quasi-identifiants sont déductibles de l'examen du jeu de données.

régression est utilisée sous la forme d'apprentissage supervisé afin de prédire la valeur d'un critère d'utilité ou d'un critère de sécurité, compte tenu de données expérimentales disponibles dans OPAM.

Enfin, l'approche s'appuie sur l'ontologie OPAM (BenFredj *et al.*, 2015). Enfin, avant de présenter les différentes étapes de MAGGO, pour faciliter la compréhension de l'approche, nous rappelons à la figure 3 les éléments principaux du méta-modèle d'OPAM. Les classes sur fond jaune sont celles qui permettent la représentation de la connaissance « théorique » relative aux techniques et algorithmes d'anonymisation. Les classes sur fond bleu permettent de décrire les concepts que nous avons définis pour décrire le processus d'anonymisation. Enfin, les classes sur fond blanc sont le support de la représentation de la connaissance empirique que les expérimentations décrites dans les articles de recherche ou accumulées au cours de nos tests des outils ont permis de constituer.

4. Description des étapes de la méthode MAGGO

Dans cette section, nous décrivons chacune des cinq étapes de la démarche schématisée à la figure 1.

4.1. Etape 1 - Chargement et qualification du contexte de l'anonymisation

Une anonymisation vise la prévention contre des attaques potentielles portant atteinte à la vie privée. Sa mise en œuvre nécessite la sélection d'une ou plusieurs techniques qui mettent en œuvre le modèle de protection censé contrer ces attaques. Ainsi se pose le problème de choix d'algorithmes pour mettre en œuvre l'anonymisation qui répond aux attentes de son initiateur. Ces attentes constituent l'ensemble des exigences que doit satisfaire l'anonymisation. A ce titre, on peut considérer deux catégories d'exigences pour l'anonymisation de micro-données. La première catégorie rassemble les exigences indépendantes de la technique par exemple l'usage prévu des données anonymes (publication, test, classification, etc.) le seuil de risque de ré-identification toléré, le taux de suppression à ne pas dépasser ainsi que la qualité minimale exigée. Cette dernière peut être exprimée par l'importance relative accordée aux critères de qualité que doivent vérifier les données anonymes. La deuxième catégorie regroupe des exigences dépendantes de la technique choisie (ici la généralisation) et influent sur le choix d'un algorithme implémentant cette technique. Dans le cas de la généralisation, le type de généralisation souhaité peut constituer une exigence spécifique. A titre d'exemple, une anonymisation par généralisation pourrait être demandée pour un besoin de classification des données, tout en exigeant de ne pas accepter un taux de suppression de plus de 5% (qui réduit l'échantillon et éventuellement le déforme) ni un résultat qui engendre un risque de ré-identification de plus de 10%. Le demandeur pourrait aussi préciser qu'il accorderait plus d'importance à la sécurité qu'à la complétude des données anonymes (dans ce cas, il exprime une préférence pour la suppression qui permet d'effacer des « outliers », présentant un risque élevé de ré-identification). Quand bien même on dispose de ces informations, elles ne

suffisent pas pour sélectionner des algorithmes adéquats. En effet, comme on a pu le constater dans notre état de l'art sur l'anonymisation par généralisation (BenFredj *et al.*, 2014), le choix des algorithmes repose sur des données descriptives de la base qui, si elles ne peuvent pas être déduites automatiquement, doivent être fournies par le demandeur. A cet effet, on peut citer la qualification des attributs (identifiant/quasi-identifiant/sensible/non sensible, catégoriel/continu). De plus, certaines données descriptives (par exemple, la liste des attributs formant le quasi-identifiant) sont nécessaires quelle que soit la technique. D'autres (par exemple, la distribution des données) sont spécifiques à une technique. En résumé, dans un souci de généricité, le contexte d'une anonymisation sollicitée par un utilisateur pour ses micro-données est construit en deux temps (Fig. 4). Dans un premier temps, MAGGO construit le contexte à qualifier en récupérant de l'ontologie ses paramètres, c'est-à-dire les types d'exigences utilisateur à renseigner ainsi que les types de données descriptives à connaître pour le type d'anonymisation sollicitée.

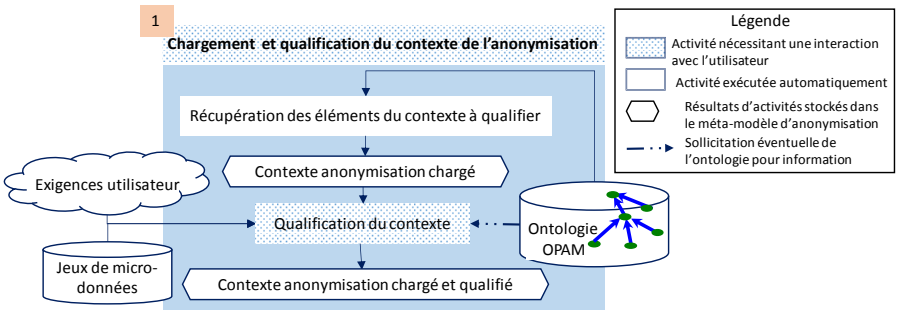


Figure 4. Chargement et qualification du contexte de l'anonymisation

Tableau 2. Paramètres de contexte de la généralisation de micro-données

Paramètres fournis par l'utilisateur	Paramètres pouvant être déduits automatiquement
Seuil de risque toléré	Attributs du QI
Taux de suppression autorisé	Attributs identifiants
Besoin d'usage	Attributs sensibles
Jeu de micro-données original	Nature de chaque attribut du QI : catégoriel ou continu
Propriétés de qualité attendues	Type de généralisation attendu
Importance relative des propriétés de	Distribution des données
	Taille du jeu de micro-données
	k : taille maximale des classes d'équivalence de QI
	MaxSup : le nombre maximal de tuples à supprimer

Certains de ces paramètres, rappelons-le, sont spécifiques à une technique. A titre d'exemple, dans le cas d'une anonymisation par généralisation, notre approche MAGGO, après interrogation de l'ontologie OPAM, construira le contexte d'anonymisation par généralisation. Ce contexte est constitué des paramètres de contexte décrits dans le tableau 2. Ces paramètres de contexte, intégrés dans le méta-modèle d'anonymisation, seront renseignés, dans la seconde phase de l'étape « chargement et qualification du contexte ». La plupart des paramètres sont déductibles de l'analyse des jeux de données. Deux paramètres spécifiques à la généralisation, MaxSup et k, sont calculés. MaxSup définit le nombre maximum de

lignes qui pourront être supprimées pendant l'anonymisation. L'attribut k fait référence au k-anonymat (Fung *et al*., 2010), modèle de protection de la vie privée ciblé par la technique de généralisation. Il correspond à la taille minimale des classes d'équivalence de quasi-identifiants anonymes pouvant être générés par généralisation. Par exemple, si le sexe et le code postal forment un quasi-identifiant et que k vaut 10, le jeu de données anonymisées ne pourra pas comprendre moins de 10 lignes pour le même sexe et le même code postal. Si nécessaire, soit le code postal sera généralisé au numéro du département soit les lignes correspondantes seront supprimées.

Ainsi, dans MAGGO, MaxSup est calculé à partir de la taille du jeu de données et du taux de suppression autorisé par l'utilisateur en appliquant la formule suivante : *MaxSup = Taille du jeu de micro-données * taux de suppression autorisé*

Pour calculer k, nous utilisons la formule suivante de l'outil PARAT :

k = 100/ taux de risque de ré-identification

Cette formule exprime le fait que le taux de risque de ré-identification est inversement proportionnel à k. En d'autres termes, plus k est petit, plus le risque de ré-identification est grand.

Une fois le contexte d'anonymisation renseigné, MAGGO suggère à l'utilisateur, dans sa seconde étape, sous forme de signatures, un ensemble potentiel d'algorithmes paramétrés susceptibles de satisfaire à ses exigences.

4.2. Etape 2 - Suggestion de signatures d'algorithmes candidats

Le jeu de données brouillé renvoyé par application d'une technique d'anonymisation dépend fortement de la signature de l'algorithme exécuté sur le jeu de données original. La construction, l'évaluation et la proposition à l'utilisateur, de signatures d'algorithmes se rapprochant le plus de ses exigences de qualité, est l'objet de cette étape de MAGGO (Fig. 5). La première phase de cette étape consiste à construire des signatures pertinentes. Dans un premier temps, on extrait les algorithmes applicables au contexte de l'anonymisation et on les dote de valeurs de paramètres conformes aux contraintes spécifiées dans le contexte. La seconde phase a pour objectif de proposer à l'utilisateur, parmi les signatures pertinentes, celles offrant le meilleur score en termes de concordance avec les exigences de qualité. Les paragraphes qui suivent détaillent chacune de ces phases.

4.2.1 Construction des signatures pertinentes

Le ciblage des algorithmes applicables au contexte exploite certains paramètres de contexte. A titre d'exemple, pour une anonymisation par généralisation, si l'utilisateur n'a pas d'exigence sur le type de généralisation à obtenir alors, de ce point de vue, tous les algorithmes de généralisation sont candidats au paramétrage.

En revanche, si son souhait est d'obtenir des généralisations multidimensionnelles[7], alors cet ensemble se restreint aux algorithmes fournissant ce type de généralisation tel que le « Median Mondrian ».

Pour effectuer ce filtrage d'algorithmes, l'ontologie OPAM est exploitée car elle dispose des connaissances permettant de confronter les exigences des algorithmes aux exigences de l'anonymisation. Ces connaissances sont celles se trouvant dans le schéma d'OPAM représenté à la figure 3.

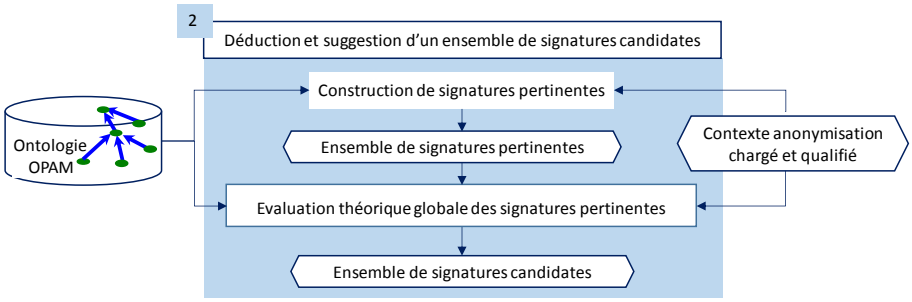


Figure 5. Déduction et suggestion de signatures candidates

Les algorithmes sélectionnés permettent bien sûr d'instancier le méta-modèle de l'anonymisation (les classes jaunes du méta-modèle de la figure 2). Cette instanciation contient aussi, pour chaque algorithme sélectionné, l'ensemble des combinaisons possibles de valeurs de paramètres pouvant lui être affectées. Chaque algorithme sélectionné couplé avec chaque combinaison de valeurs de paramètres possible constitue une signature pertinente.

Il s'agit d'octroyer au paramètre de l'algorithme, la valeur de contexte générée suite à la prise en compte de la contrainte d'anonymisation imposée par l'utilisateur. A titre d'exemple, dans le cas d'une anonymisation par généralisation, l'utilisateur exprime un taux de risque de ré-identification et un taux de suppression tolérés (paramètres saisis). Ces deux contraintes génèrent dans le contexte de l'anonymisation une valeur pour k et MaxSup (paramètres générés). Ces deux valeurs, combinées avec chaque algorithme retenu, constituent autant de signatures.

4.2.2. Evaluation théorique des signatures pertinentes

Cette phase vise à fournir à l'utilisateur les signatures se rapprochant le plus de ses exigences de qualité et de sécurité. C'est un processus de décision multicritère pour lequel nous appliquons la méthode AHP. Cette dernière, sur la base des comparaisons par paires, détermine le score global de chacune des signatures afin de retenir les mieux classées. On peut ainsi décider de fournir à l'utilisateur les trois signatures pertinentes ayant le score le plus élevé.

[7] Une généralisation multidimensionnelle est telle que, dans la table résultat, les données ne sont pas nécessairement au même niveau de généralité. Ainsi, on peut imaginer qu'une tranche d'âge pourra être plus ou moins large selon les individus.

La hiérarchie fournie à AHP a pour premier niveau l'objectif de cette étape. Le niveau intermédiaire correspond à la hiérarchie des exigences emmagasinée dans OPAM. Son dernier niveau, c'est-à-dire les feuilles de l'arbre, regroupe les signatures pertinentes à évaluer. La figure 6 contient, à titre d'exemple, la hiérarchie construite par cette phase pour une anonymisation à des fins de classification.

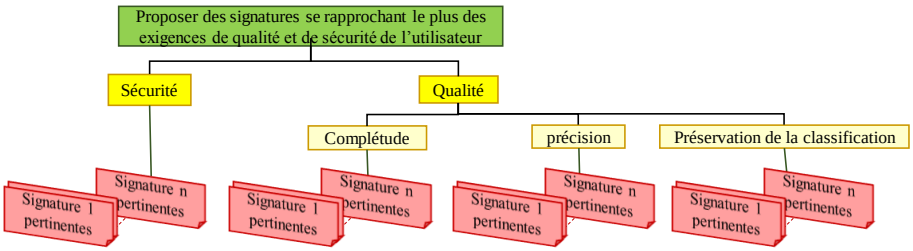


Figure 6. Hiérarchie multicritère pour l'anonymisation

Une fois la hiérarchie construite, les jugements sur l'importance relative des éléments de cette hiérarchie sont déterminés. Les jugements entre les éléments du niveau intermédiaire de la hiérarchie (les critères) sont ceux émis par l'utilisateur et spécifiés dans le contexte de l'anonymisation. Les jugements sur l'importance relative des signatures sont, quant à eux, déterminés de façon automatique après une évaluation de chaque signature selon un critère donné. Cette évaluation approximative, que l'on nomme « évaluation théorique locale », est déduite des expérimentations faites par les experts en anonymisation et qui sont emmagasinées dans OPAM (classes sur fond blanc de la figure 3). L'importance relative de chaque signature est aussi déterminée automatiquement. Elle est fondée sur leur évaluation locale et sur une échelle de comparaison disponible dans MAGGO.

Les paragraphes qui suivent décrivent respectivement les processus d'évaluation locale et globale (le score) d'une signature.

4.2.2.1. Evaluation théorique locale des signatures pertinentes

Plusieurs évaluations théoriques d'algorithmes d'anonymisation de micro-données sont disponibles dans la littérature. Chacune fournit la qualité d'un jeu de données anonyme vis-à-vis d'un critère (sécurité, précision, complétude, etc.) compte tenu d'une signature d'algorithme et des caractéristiques spécifiques du jeu de données originales. Le critère en question est mesuré à l'aide d'une métrique. Dans le cas où il n'y a pas d'évaluation théorique, pour la signature et les caractéristiques du jeu de données spécifiées dans le contexte d'anonymisation, une technique d'apprentissage supervisée est mise en place afin de prédire la qualité de cette signature vis-à-vis d'un critère. La régression se prête bien à notre problématique en raison du type des variables explicatives et de la variable cible. Le modèle retenu est l'arbre de régression en raison de la petite taille de la base d'expérimentations disponibles (Loh, 2011). La variable à expliquer est le critère de qualité à mesurer. Les variables explicatives sont les différents éléments de contexte

influençant la variable cible. Le jeu de données d'entrainement est extrait de l'ontologie OPAM. Un exemple est constitué d'une entrée et d'une sortie.

Ainsi, à titre d'exemple, pour une anonymisation par généralisation à des fins de classification, il nous faut quatre jeux de données : un par critère constituant une feuille du niveau intermédiaire de la hiérarchie AHP (sécurité, complétude, précision, préservation de la classification) décrite à la figure 6. Tous les jeux de données contiennent les mêmes informations : une valeur pour « k », une valeur pour « nombre d'attributs constituant le QI », une valeur pour « distribution du jeu de micro-données original ». En revanche, ces jeux d'exemples se distinguent par la sortie qui correspond à la mesure du critère cible.

Après évaluation de chaque signature, le méta-modèle est enrichi de ces nouvelles estimations.

4.2.2.2. Mesure de l'importance relative des signatures

Tableau 3. Comparaison des signatures par une échelle sémantique

Intensité	Signification	Interprétation formelle de la signification
(Sj, Sj', 1)	Sj et Sj' sont d'égale qualité vis-à-vis du critère Ci	$E_{Sj}^{Ci} - E_{Sj'}^{Ci} \leq \varepsilon_1$
(Sj, Sj', 2)	Sj est d'une qualité légèrement meilleure que celle de Sj' vis-à-vis du critère Ci	$\varepsilon_1 < E_{Sj}^{Ci} - E_{Sj'}^{Ci} \leq \varepsilon_2$
(Sj, Sj', 3)	Sj est d'une qualité meilleure que celle de Sj' vis-à-vis du critère Ci	$\varepsilon_2 < E_{Sj}^{Ci} - E_{Sj'}^{Ci} \leq \varepsilon_3$
(Sj, Sj', 4)	Sj est d'une qualité nettement meilleure que celle de Sj' vis-à-vis du critère Ci	$\varepsilon_3 < E_{Sj}^{Ci} - E_{Sj'}^{Ci} \leq \varepsilon_4$
(Sj, Sj', 5)	Sj est d'une qualité très nettement meilleure que celle de Sj' vis-à-vis du critère Ci	$\varepsilon_4 < E_{Sj}^{Ci} - E_{Sj'}^{Ci} \leq \varepsilon_5$

Une fois les évaluations locales des différentes signatures effectuées, il s'agit de procéder à des comparaisons par paires de signatures afin de déduire l'importance relative des signatures vis-à-vis de chaque critère. Pour ce faire, nous nous inspirons de l'échelle sémantique de (Saaty et Sodenkamp, 2008) afin de permettre une comparaison automatique par paires de signatures pour livrer à AHP la matrice de comparaison des signatures pertinentes. Si l'on considère deux couples E(Ci,Sj) et E(Ci,Sj') où E(Ci, Sj) (resp. E(Ci,Sj')) représente l'évaluation locale de la signature Sj (resp. Sj') pour le critère Ci, nous construisons la table d'échelle sémantique d'AHP comme suit (Tableau 3). Dans cette table servant de comparaison par paires de signatures $\varepsilon1 < \varepsilon2 < \varepsilon3 < \varepsilon4 < \varepsilon5$. Ces valeurs sont définies par l'approche pour chaque critère de qualité.

4.3. Etapes 3,4 et 5 de MAGGO

Une fois la comparaison par paires effectuée, AHP se charge de fournir le score global de chaque signature pertinente ; ce qui permet de classer ces signatures et de proposer à l'utilisateur, dans l'étape 3 de MAGGO, les signatures qui ont le meilleur score. Ce dernier a la possibilité de choisir une ou plusieurs signatures à faire

exécuter sur son jeu de micro-données. L'exécution de ces signatures fait l'objet de l'étape 4 de MAGGO. Dans cette étape, un jeu de données anonyme est livré pour chaque signatures pertinentes, de score le plus élevé, choisies par l'utilisateur. Pour guider l'utilisateur dans son choix de jeu de données anonymes, différentes évaluations cette fois-ci réelles, sont effectuées. Chaque évaluation permet de positionner le jeu anonyme vis-à-vis d'une exigence de qualité attendue.

5. Exemple d'illustration

Pour illustrer notre approche, on suppose que le contexte est caractérisé comme suit. Le risque maximum toléré est 10%. De même, on admet que l'on ne peut supprimer plus de 20% des données. De plus, la table à anonymiser est de grande taille (1000 tuples). La distribution des données est dense. Le quasi-identifiant comprend trois attributs. L'usage des données anonymisées est la classification. L'utilisateur accorde autant d'importance à l'utilité des données qu'au respect de la vie privée.

Etape 1 – chargement et qualification du contexte

Au cours de cette première étape, l'éditeur de données doit entrer son contexte. Certains éléments (taille, distribution, nombre d'attributs du QI) peuvent être calculés automatiquement après chargement de la table.

Etape 2 – Sélection d'algorithmes et signatures pertinentes

Les paramètres k et MaxSup peuvent être calculés en fonction du taux de risque et du taux de suppression. Ici k vaut donc 10 et MaxSup=20*1000/100=200. Plus précisément 10 est la valeur minimale de k et 200 la valeur maximale de MaxSup. On peut aussi tester des signatures où k=12 et MaxSup=150 par exemple.

L'algorithme de Samarati (2001) ne peut pas être appliqué à une table de cette taille, car il est trop gourmand en temps de réponse. Cette information fait partie des connaissances contenues dans l'ontologie. Supposons donc que seuls les algorithmes Datafly, Median Mondrian et TDS remplissent les contraintes.

Les deux phases précédentes de l'étape 2 ont généré deux valeurs de k (10 et 12), deux valeurs de MaxSup (200 et 150) et trois algorithmes (Datafly, Media Mondrian et TDS). Seul Datafly effectue des suppressions. Par conséquent, les signatures générées sont récapitulées dans les quatre premières colonnes du tableau 4. Elles sont évaluées selon les critères feuilles de la hiérarchie des buts (Fig. 6). Les évaluations liées aux deux critères 'sécurité' et 'complétude' ont été déduites à partir respectivement des valeurs de k et MaxSup. Celles liées aux critères 'Précision' (métrique de discernabilité DM (Fung *et al.,* 2010)) et 'Préservation de la classification' ont été déduites en appliquant la régression sur les données expérimentales issues d'OPAM (Tableau. 4).

Le passage des évaluations individuelles des signatures à des comparaisons deux à deux est nécessaire afin de pouvoir appliquer la méthode AHP. Par exemple, pour le critère classification, les signatures 5 et 8 sont évaluées respectivement à 0,65 et

0,71, ce qui représente une différence de 6%. On suppose que l'échelle utilisée induit ainsi une intensité de 3. Les huit signatures sont ainsi comparées deux à deux pour chacun des critères. On aboutit à un score final fourni en dernière colonne du tableau 4. Après application d'AHP, il agrège les quatre critères pour chaque signature. Ce score permet à l'utilisateur de choisir d'exécuter les signatures (par exemple les quatre dernières) qui donnent le meilleur compromis entre les quatre critères, compromis qui résulte de l'application d'AHP à chaque paire de signatures.

Tableau 4. Evaluation locale (individuelle) des signatures

Signature	Algorithme	k	Maxsup	Sécurité	Complétude	Précision métrique DM	Usage Classification	Score final
Sig 1	Datafly	10	150	0,9	0,85	50000	0,54	0,1
Sig 2	Datafly	10	150	0,9	0,85	50000	0,54	0,05
Sig 3	Datafly	12	200	0,92	0,8	60000	0,61	0,04
Sig 4	Datafly	12	200	0,92	0,8	60000	0,61	0,05
Sig 5	Mondrian	10	0	0,9	1	15000	0,65	0,27
Sig 6	Mondrian	12	0	0,92	1	20000	0,63	0,18
Sig 7	TDS	10	0	0,9	1	35000	0,79	0,19
Sig 8	TDS	12	0	0,92	1	40000	0,71	0,12

6. Conclusion

Les propriétaires de données sont confrontés à deux difficultés majeures lors d'un processus d'anonymisation. La première concerne le choix de l'algorithme adéquat au contexte. La seconde est le paramétrage de telle sorte qu'il délivre des données sécurisées (difficiles à ré-identifier) et utiles (dont la qualité reste conforme avec l'objectif). Notre approche MAGGO automatise ces deux tâches en utilisant une ontologie. Cette dernière peut aussi être consultée par le propriétaire des données afin de recueillir les connaissances nécessaires lui permettant de décrire son contexte et de répondre de façon adéquate aux questions qui lui sont posées lors du déroulement du processus. La sécurisation des données par anonymisation et le maintien de la précision et de la complétude des données sont contradictoires. C'est pourquoi, le processus d'anonymisation vise un compromis entre ces deux objectifs, en fonction de l'usage des données. Notre approche est, pour le moment, limitée aux algorithmes fondés sur la technique de généralisation. Toutefois, nous nous sommes efforcées de la rendre la plus générique possible afin qu'elle puisse être appliquée à d'autres techniques d'anonymisation de micro-données. Enfin, pour rendre l'approche évolutive et son implémentation incrémentale, nous avons utilisé une conception dirigée par les modèles.

En termes de recherche future, nous envisageons trois axes : 1) la mise au point d'un outil support de l'approche, 2) la conduite d'une expérimentation à plus grande échelle incluant des utilisateurs pour mesurer l'utilité et l'utilisabilité de la méthode et de l'outil, 3) l'extension à d'autres techniques pour pouvoir choisir à la fois une technique d'anonymisation, un algorithme et une signature.

Bibliographie

BenFredj F., Lammari N., Comyn-Wattiau I (2014), Characterizing Generalization Algorithms-First Guidelines for Data Publishers. International Conference on Knowledge Management and Information Sharing, Rome, Italy.

BenFredj F., Lammari N., Comyn-Wattiau I. (2015) Building an Ontology to Capitalize and Share Knowledge on Anonymization Techniques. *European Conference on Knowledge Management: 122-131*. Kidmore End: Academic Conferences International Limited.

Brand, R, (2002) Microdata protection through noise addition. In Domingo-Ferrer J, editor, *Inference Control in Statistical Databases*, LNCS Vol. 2316, pp 97-116, Springer.

Dai C, Ghinita G, Bertino E, Byun J, Li N (2009) TIAMAT: a Tool for Interactive Analysis of Microdata Anonymization Techniques. PVLDB 2(2): 1618-1621 (2009)

Defays, D., Nanopoulos, P. (1993) Panels of enterprises and confidentiality: the small aggregates method, Paper read at the 92nd Symposium on Design and Analysis of Longitudinal Surveys, Ontorio, Canada, November.

Fung, B. C. M., Wang, K., Chen, R., Yu, P. S. (2010) Privacy preserving data publishing: a survey of recent developments. In ACM Computing Surveys (CSUR), Vol. 42(14).

Hand D.J., 1992. Microdata, macrodata, and metadata. In Dodge Y., Wittaker J. (Eds), *Computational Statistics*, Physica Verlag, Heidelberg, p 325-340.

Fienberg S.E, McIntyre J. (2004). Data swapping: Variations on a theme by dalenius and reiss. In International Workshop on Privacy in Statistical Databases (pp. 14-29). Springer

Ilavarasi. B., Sathiyabhama A. K., Poorani. S. (2013) A survey on privacy preserving data mining techniques. *Int. Journal of Computer Science and Business Informatics*, 7(1).

Loh W-Y (2011) Classification and regression trees. Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery 1(1): 14-23.

Patel, L., Gupta, R. (2013) A Survey of Perturbation Technique for Privacy-Preserving of Data. In Int. Journal of Emerging Technology and Advanced Engineering, Vol 3(6).

Poulis G., Gkoulalas-Divanis A., Loukides G., Skiadopoulos S., Tryfonopoulos C.:SECRETA: A System for Evaluating and Comparing RElational and Transaction Anonymization algorithms. EDBT 2014.

Saaty T.L, Sodenkamp M.A. (2008) Making decisions in hierarchic and network systems. IJADS 1(1): 24-79

Samarati, P. (2001) Protecting respondents' identities in microdata release, *IEEE Transactions on Knowledge and Data Engineering*, Vol 13, No. 6, pp 1010-1027.

Silver M. S. (2006) Decisional Guidance. Broadening the Scope. In: Galleta, D. and Zhang, P. (eds.). Human-Computer Interaction in Management Information Systems. *International handbooks on information systems* Vol 6, pp 90-119. Armonk, NY: M.E. Sharp.

Xiao X, Wang G, Gehrke G, (2009) Interactive Anonymization of Sensitive Data. SIGMOD'09, June 29–July 2, 2009, Providence, Rhode Island, USA, pages 1051–1054.

Modéliser l'avion et son moyen de production : vers un modèle global pour de la conception simultanée

François Bouissiere[1], Claude Cuiller[1], Pierre-Eric Dereux[1], Stéphane Kersuzan[1], Thomas Polacsek[2]

[1] *AIRBUS, 1 Rond Point Maurice Bellonte,* 31707 *Blagnac, France*
[2] *ONERA, 2 avenue Edouard Belin BP74025, 31055 TOULOUSE Cedex 4*

RÉSUMÉ. *La construction d'objets complexes, tel qu'un avion, nécessite souvent la création d'un système industriel dédié. Par système industriel, nous entendons l'ensemble des moyens, matériels et immatériels, utilisés pour construire l'objet (main d'œuvre, machines, etc.). Classiquement, ce moyen de production est défini après la conception du produit. En d'autres termes, les spécifications du produit permettent de définir les exigences du système industriel. Le défaut de cette approche est que le système industriel peut hériter de contraintes bloquantes qui pourraient être aisément levées en changeant la conception du produit. Par conséquent, il est intéressant de définir un cadre global permettant de concevoir à la fois le produit et son système industriel. Cet article présente un cas d'étude industriel, dans le domaine aéronautique, qui peut servir de support à l'élaboration d'un tel cadre et propose un premier prototype de modèle pour une conception simultanée de l'avion et de sa production.*

ABSTRACT. *The construction of complex objects, such as an aircraft, requires the creation of a dedicated industrial system. By industrial system, we mean all the material and immaterial means used to build the object (labor, machines, factories, etc.). Classically, this means of production is defined after that the design of the product. In other words, the specifications of the product are the requirements of the industrial system. The drawback of this approach is that the industrial system can inherit blocking constraints that could be easily removed by changing the design of the product. Therefore, it is useful to define a global framework for designing both the product and its industrial system. This article presents an industrial case study, in the field of aeronautics, which can be used to support the development of such framework and proposes a first prototype model for a simultaneous design of the aircraft and its production.*

MOTS-CLÉS : *Ingénierie dirigée par les modèles, conception simultanée produit production*

KEYWORDS: *Model-driven engineering, simultaneous conception product production*

1. Introduction

Aujourd'hui, le marché aéronautique évolue extrêmement vite, créant ainsi une désynchronisation entre, d'un côté, les cycles de développement et de vie d'un avion et, de l'autre côté, les évolutions rapides du marché. En particulier, l'émergence régulière de compagnies aériennes porteuses de nouveaux modèles d'affaire contraignent les constructeurs aéronautiques à rechercher des solutions innovantes pour permettre de réduire drastiquement la durée des cycles de développement et les coûts de production. Ces nouveaux clients demandent des alignements très rapides entre les modèles d'avions et leurs attentes, que cela soit en coûts d'exploitation comme en capacité d'évolution de l'avion durant son exploitation. La réactivité exigée ici rentre directement en conflit avec le cycle de vie des avions qui s'inscrivent eux dans des temps longs.

Par conséquent, les constructeurs aéronautiques sont à la recherche de nouvelles solutions, permettant de réduire les cycles de conception et de production tout en limitant, le plus possible, une augmentation des coûts. De par la nature industrielle du problème, le retour sur investissement est un paramètre clef de toute adoption de nouvelles méthodes. En outre, dans un marché globalisé, l'arrivée de nouveaux constructeurs oblige les constructeurs historiques à introduire plus d'innovation dans leurs produits, et dans les méthodes de conception, pour conserver leur avantage compétitif.

C'est justement à la confluence de ces problèmes que se trouve l'*architecte avion*. Dans l'industrie aéronautique, l'architecte avion a pour rôle de gérer les interactions entre les différentes composantes de l'avion, qui s'inscrivent dans des disciplines très variées (aérodynamique, sécurité, thermique, etc.), afin de répondre aux mieux aux exigences aussi bien de l'avion que de fabrication.

Ainsi, la définition d'un avion et de son système de production[1] est le résultat d'activités itératives de conception, où des compromis sont sans cesse effectués pour converger sur un optimum global satisfaisant à la fois les intérêts du constructeur et ceux des compagnies aériennes. Pour réaliser ces arbitrages, il est nécessaire, pour l'architecte avion, de disposer d'une vue globale et cohérente des principaux composants de l'avion et, notamment, des interactions avec le moyen de production. Or, si la partie conception et la partie industrialisation sont aujourd'hui bien maîtrisées séparément, leurs interactions restent excessivement complexes à appréhender.

Nous pensons qu'une approche Ingénierie Systèmes pourrait permettre de passer un cap, en proposant une vision plus intégrée de ses deux versants du métier d'architecte. Plus précisément, disposer de modèles conceptuels permettrait de poser clairement, et de comprendre, pour chacun des acteurs impliqués, les problèmes sous jacents à la création d'une vision globale du produit et de son système industriel dans le cadre aéronautique. De plus, même si nous n'adressons pas le problème dans cet ar-

1. Par système de production nous entendons l'ensemble des moyens, matériels et immatériels, utilisés pour construire un avion (main d'œuvre, machines, usines, méthodes, outils, etc.).

ticle, des travaux ont montré la possibilité d'utiliser des solveurs logiques pour l'aide à la conception (Delmas *et al.*, 2011 ; Delmas et Polacsek, 2013). Dans la lignée de ces travaux, nous avons pour objectif, dans les travaux futurs, de faire appel à des méthodes automatiques pour pouvoir optimiser de concert la conception et l'outil industriel. Pour cela, il sera nécessaire de disposer d'un double digital de l'usine et de l'avion, double qui permettra de réaliser les opérations d'analyses et d'optimisations conjointent du produit et de son moyen de production.

Dans cet article, nous allons présenter une activité initiée chez Airbus pour accompagner le développement incrémental du programme d'avions A320 visant la montée en cadence de la chaîne de production. Cette augmentation de la cadence a pour but de répondre aux besoins des compagnies aériennes, tout en s'accompagnant de réduction des coûts récurrents de fabrication et de l'amélioration de la capacité à fournir de nouveaux services. Les travaux présentés ici sont encore au stade préliminaire. Ils représentent un effort important, de plusieurs mois, de collecte d'informations de nature très variée (documents, entretiens, plans, etc.) et de données brutes dans le cadre d'un cas d'étude concret, qui ont permis de faire émerger un modèle préliminaire, support du futur double digital.

Notre but ici est donc de présenter un problème industriel et montrer comment il a alimenté nos premières pistes de réflexions en vue de faire un cadre pour la conception simultanée, c'est-à-dire ici la conception en parallèle d'un produit et de son moyen de production. Des cadres génériques pour la modélisation conjointe de la conception et de la production existent déjà (Sprock et McGinnis, 2015) (Demoly *et al.*, 2011) (Benkamoun *et al.*, 2014), cependant, ces cadres restent de très haut niveau et son difficile à instancier sur un cas industriel concret. Nous avons ici fait la démarche inverse consistant à partir des données industrielles sur un cas existant et en faire émerger un cadre dédié à la pratique aéronautique.Nous espérons que notre cas d'étude pourra nourrir une réflexion plus large dans le futur.

Notons que le principe de conception simultanée entre le bureau d'étude et la production est un concept relativement ancien (Decreuse et Feschotte, 2017 ; Shenas et Derakhshan, 1994). Employée dans l'automobile, notamment dans le cadre de la logistique des éléments fournis par les sous-traitants (Göpfert et Schulz, 2013), sa mise en œuvre dans le cadre de l'étude de modifications de l'A320 pose des problèmes particuliers du fait de certaines spécificités du monde aéronautique. En effet, un avion est un objet particulièrement complexe composé de nombreux éléments requérant des opérations de fabrication souvent manuelles et faisant appel à un savoir-faire très spécifique. La conception, elle, est réalisée de manière séquentielle en considérant d'abord les exigences liées aux performances du produit (nombre de passager, consommation, etc.), puis la définition des composants majeurs de l'avions pour seulement finir par la définition des moyens de productions. De plus la réglementation aéronautique impose, au travers de la certification, des contraintes très fortes sur la conception, la fabrication et l'opération d'avion. Ainsi, ces contraintes mettent un frein aux évolutions des méthodes de conception et de production du fait des efforts nécessaires pour fournir les éléments justifiant du maintien de la conformité aux ré-

glementations. Enfin les volumes de production, limités à quelques centaines d'avions par an, n'ont pas incité à revoir les modes de production qui font essentiellement appel à des tâches manuelles.

Dans la Section 2 de cet article, nous présenterons notre problème et nous donnerons une description du contexte industriel. Dans la Section 3, nous donnerons les concepts clés de la chaîne de production et expliquerons la démarche que nous avons suivie pour capturer cette connaissance. Nous expliciterons et motiverons notre modèle dans la Section 4 et la Section 5 sera consacrée à la conclusion.

2. Présentation de notre problème

2.1. *Contexte industriel*

Le cadre industriel de notre étude concerne la partie de la chaîne de production qui réalise l'assemblage des *tronçons avants* pour les avions monocouloir de la famille A320. Le tronçon avant englobe la pointe de l'avion, le poste de pilotage, et s'étend dans la cabine jusqu'au tronçon central où se trouve la jonction avec la voilure. La chaîne de production du tronçon avant est située sur le site de Saint-Nazaire et s'organise au travers de deux sous chaînes, la première dédiée à l'ensemble des assemblages dit structuraux, le corps de l'avion, et la seconde, nommée COMETE, dédiée à l'installation de l'ensemble des systèmes. Ici, ce que nous appelons systèmes correspond en fait à des éléments non structuraux tels que : l'isolation thermo-phonique, les harnais et meubles électriques, les tuyauteries d'air conditionné, etc.

Dans cette étude, nous nous sommes concentré sur la chaîne COMETE (Figure 1). Cette chaîne est constituée de 14 stations sur lesquelles sont distribuées les activités manufacturières, c'est-à-dire d'installation d'équipements sur le tronçon avant. Elle est pilotée sur un principe dit de *ligne pulsée* qui consiste à faire avancer tous les X heures le tronçon d'avion à équiper et ses outillages de station en station. Le produit arrive donc sur la station 1 où débute l'installation des équipements et sort complètement équipé sur la station 14.

Cette famille d'avion étant ancienne, sa conception n'a pas été prévue pour de fortes cadences. A cette époque, la cadence d'un avion par mois était une performance remarquable. Aujourd'hui, les cadences atteignent un rythme de 50 avions par mois pour répondre à la demande du marché. Il faut réaliser que, de par le poids de l'histoire de ce programme, les activités manufacturières sont fortement manuelles et les technologies utilisées, d'une grande fiabilité et d'un coût assez faible, n'ont quasiment pas évolué depuis les origines. De plus, les zones de travail ont conservé leur exiguïté, rendant les tâches des opérateurs particulièrement pénibles aux cadences actuelles.

Notre problème industriel consiste donc à élaborer des solutions pour atteindre des cadences de 63 avions par mois (et plus), sans introduire de changement profond dans la chaîne. Ces changements doivent être à faible impact sur la chaîne de par la volonté

Figure 1. *Chaîne d'installation des systèmes COMETE*

de minimiser les risques industriels. De plus, il faut éviter la solution qui consisterait à construire une nouvelle chaîne d'installation en renfort de l'existant.

Une option possible pour adapter les cadences à la demande croissante, sans impacter le moyen de production, ou presque, est de réaliser des changements dans la conception même de l'avion pour rendre les opérations manufacturières "*plus simples*". Le véritable défi est, qu'aujourd'hui, la conception n'est pas pensée en termes d'installation. Un élément à installer peut être lourd et grand, donc difficilement manœuvrable, mais peut devoir être installé dans une position non ergonomique pour un opérateur, par exemple les bras en l'air. Si nous choisissons de déplacer cet élément dans un endroit plus accessible, ou carrément de changer cet élément, nous pouvons espérer des gains dans le temps d'installation. Pour réaliser cela, il nécessaire que l'architecte dispose d'une vue précise des conditions d'installation dès les phases de conception.

Afin de répondre à ces challenges, nous avons décidé de mettre en place une activité de conception simultanée entre le bureau d'études et la production. Le but de cette conception simultanée est de pouvoir itérer rapidement entre des propositions d'évolution de la définition de l'avion et l'évaluation des bénéfices attendus côté chaîne de production (réduction du cycle d'assemblage). Pour ce faire, il est nécessaire d'identifier et de caractériser les interactions entre la conception de l'avion et son système de production.

2.2. *Un problème d'abstraction*

La conception des avions de la famille A320 a suivi un développement très traditionnel, en cascade, partant d'exigences de très haut niveau raffinées en exigences de plus bas niveau et traduites en spécifications. Ces étapes de conception datant, peu d'outils informatiques ont été utilisés que cela soit pour l'ingénierie comme pour la production.

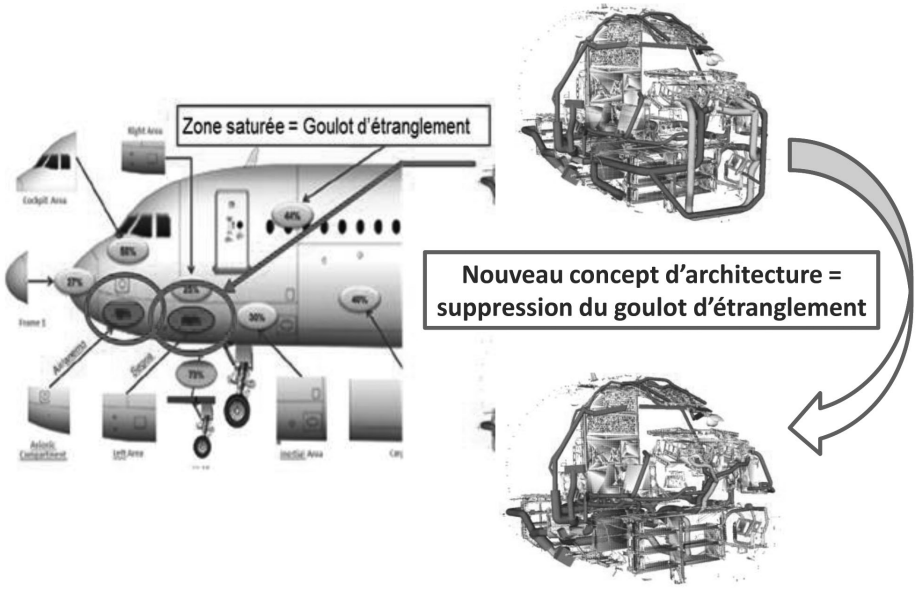


Figure 2. *Concept d'architecture permettant de supprimer un goulot d'étranglement*

Dans cette approche séquentielle, en cascade, les phases d'avant-projet se consacrent exclusivement à la spécification générale et au dimensionnement global de l'avion. Elles consistent en une concertation entre les architectes avion, les bureaux d'études et les entreprises nationales européennes qui constituent le *Groupement d'Intérêt Economique* (GIE) Airbus. Suite à cette étape, vient la phase d'études générales qui, à partir de l'architecture du nouvel avion, définit les principaux lots à développer par les différentes entreprises du GIE. A partir de là, débute la conception avec la définition détaillée des constituants majeurs (fuselage, poste de pilotage, ailes, empennages, systèmes, trains d'atterrissage, interface moteur). C'est sur la base de cette conception détaillée de l'avion que commencent les phases de définition des moyens de production industrielle. Par conséquent, la conception des moyens de fabrication est complètement désynchronisée par rapport aux activités d'ingénieries.

Avec cette approche, la conception et la définition de l'avion sont figées avant celles du système de production. En outre, l'organisation industrielle impose une répartition des activités prédéfinie, non modifiable (car contractuelle), laissant peu de place à l'optimisation du produit ou du système industriel.

Concernant la définition des moyens de production industrielle, pour définir les étapes d'assemblage, identifier les pièces à usiner et à assembler, les outillages, quantifier les besoins en personnel de fabrication et ordonnancer la production, il est actuellement nécessaire de disposer d'une définition très précise de l'avion. Or, la piste d'amélioration privilégiée dans cette étude consiste à pouvoir anticiper, dès les pre-

mières étapes de conception, l'impact d'une architecture avion, d'une conception d'avion, sur la chaîne de production.

Prenons l'exemple concret suivant. Après étude sur la chaîne d'assemblage, il apparaît qu'un goulot d'étranglement pour l'augmentation de la cadence est liée à l'installation d'une conduite de conditionnement d'air (voir Figure 2). Une partie de cette conduite passe dans une zone très exiguë, ce qui rend l'installation particulièrement difficile. Les architectes proposent de modifier la "*route*" de cette conduite, son cheminement dans l'avion, et de la faire passer par une zone plus facile d'accès pour les opérateurs. Pour connaitre son impact sur la chaîne de production, il est nécessaire aujourd'hui de repasser par le cycle complet : avant-projet, études générales, conception détaillée et définition industrielle.

En d'autres termes, une amélioration de l'architecture de l'avion visant à augmenter les cadences doit passer par :

1) l'identification des goulots d'étranglements en production et de leurs causes ;

2) l'identification et l'analyse des instructions d'assemblage pour construire les liens entre opérations d'assemblage, durée, pièces, zones avion, nombre d'opérateurs ;

3) l'élaboration d'un nouveau concept d'architecture visant à supprimer le goulot d'étranglement (saturation de la zone de travail, taches trop complexes, etc.) ;

4) l'analyse détaillée par l'équipe production des impacts sur les instructions, la création d'un nouveau séquencement et le calcul de la nouvelle cadence de ligne (fait principalement à la main), pour valider ou non les améliorations ciblées sur le cycle de production.

Ce processus est trop long par rapport à la réactivité exigée aujourd'hui. Alors que le travail de conception de l'avion commence avec des concepts généraux correspondant à des niveaux d'abstraction supérieurs il est nécessaire, pour connaître l'impact sur le moyen de production, de disposer de la définition du produit final. En effet, l'architecte considère des éléments comme une aile ou un train d'atterrissage alors que la fabrication manipule vis, écrou, câbles, etc. Ce processus, en plus d'être consommateur de temps, peut demander un travail qui s'avère parfois inutile quand la modification n'a pas les effets positifs escomptés sur la chaîne. Dans ce cas-là, le travail effectué est pure perte.

Pour pouvoir atteindre notre objectif de faire de la conception simultanée, nous devons pouvoir faire cette évaluation d'impact beaucoup plus rapidement, en restant au bon niveau de granularité. Il faut donc créer des niveaux d'abstraction du moyen de production afin de pouvoir calculer l'impact sur la production au même niveau d'abstraction que celui utilisé dans la conception.

2.3. *Notre objectif*

Comme nous l'avons vu, de par les raffinements successifs nécessaires, il est excessivement coûteux d'évaluer l'impact de nouvelles propositions d'architecture sur la

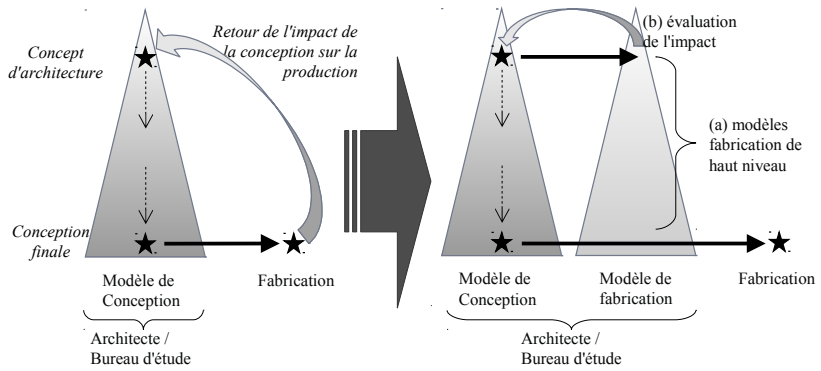


Figure 3. *Vers un modèle enrichi pour l'architecte avion*

production. Pour réduire les temps, et les coûts, de ce calcul d'impact il est nécessaire de recourir à des modèles d'abstraction de l'avion et de son système de production (voir Figure 3). De plus, l'ensemble des données manipulées aujourd'hui sont d'une granularité extrêmement fine, rendant impossible une analyse à un au niveau d'abstraction, comme c'est le cas pour un proposition de concept d'architecture. Il est donc nécessaire d'agréger ces données en des ensembles plus abstrait. Par conséquent, nous visons à terme les objectifs suivants :

(a) pouvoir raisonner au même niveau d'abstraction entre conception et fabrication. Pour ce faire, nous devons disposer d'un moyen permettant de traduire une proposition de conception en objets d'un modèle, de haut niveau d'abstraction, faisant la jonction entre conception et fabrication. Nous devons donc disposer d'abstractions de haut niveau du modèle de production au même titre que les modèles produit utilisés en phase de conception ;

(b) pouvoir faire de l'estimation de performance du moyen de production, en d'autres termes, pouvoir calculer une estimation de la cadence. Sur le modèle de (Delmas *et al.*, 2011), nous voulons pouvoir calculer l'enchaînement optimal des actions, toujours à haut niveau d'abstraction, nécessaire à la production pour en déduire une cadence de production. Les modèles définis en (a) doivent donc pouvoir servir de support à de la simulation et, plus précisément, à un outil de recherche opérationnelle qui permettra de faire l'ordonnancement (Graham, 1966).

Notre objectif à long terme est de disposer d'une chaîne complète de méthodes et d'outils permettant d'évaluer les proposions d'architecture en terme d'impact sur la production. Notons que cette approche ne se bornerait pas à calculer les séquences de montage de l'avion, mais pourrait, in fine, déduire le dimensionnement idéal de l'outil de production. Pour le moment, nous ne nous focalisons que sur la partie (a), c'est à dire sur l'élaboration de modèles de fabrication de haut niveau et leurs correspondances avec les modèles de conception.

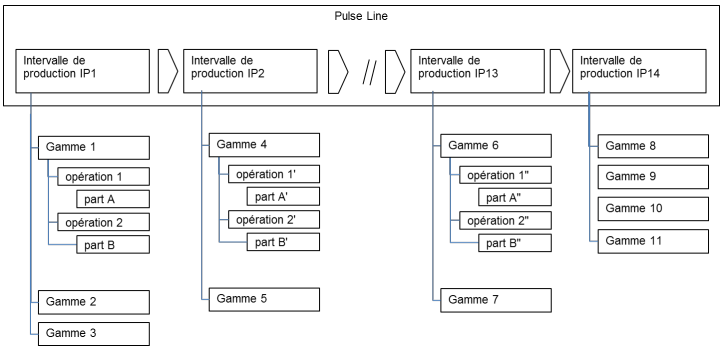


Figure 4. *Structure de la chaîne d'installation COMETE*

3. La production : un monde d'opérations

La modélisation du moyen de production nécessite une compréhension détaillée des règles qui le régissent. La chaîne de production fait intervenir de nombreuses compétences et de nombreux métiers, parfois très éloignés du monde de la conception. Compte tenu de l'ancienneté du programme A320, nous nous retrouvons avec une documentation peu digitalisée, un savoir humain considérable qui n'est parfois partagé qu'oralement, voire qui n'est qu'implicite. Par conséquent, il nous a fallu aller régulièrement sur le site de Saint-Nazaire pour interviewer les différents acteurs, comprendre et questionner les pratiques et interpréter les informations de production.

Concernant la compréhension de la chaîne de production et nos besoins, nous avons pu isoler six grands concepts (voir Figure 4) :
Pièces : éléments physiques manipulés par l'opérateur et montés sur l'avion.
Opération SOI[2] : liste les pièces à manipuler et à installer sur avion. C'est une liste des tâches à exécuter par l'opérateur, dans une séquence imposée, pour installer des pièces avion.
Gamme : regroupement d'opérations relatives à un ensemble de pièces interconnectées. La gamme décompose le travail à faire (par exemple "*installer un harnais électrique dans le poste pilotage*") en opérations (telles que : positionner le harnais derrière la console, fixer un support, connecter le harnais au calculateur, etc.).
Intervalle de Production (IP) : espace temporel, correspondant au temps qui cadence la chaîne, pendant lequel un certain nombre de gammes de montage sont à exécuter. Par abus de langage, l'IP est aussi utilisé pour désigner la station, l'espace physique sur lequel est posé le tronçon de l'avion et sur lequel sont organisés tous les éléments nécessaires, outils et pièces. Ces éléments sont listés dans les gammes associées à l'IP.
Ligne pulsée (pulse line) : composée de stations (emplacements physiques correspondant à chaque IP), elle livre les tronçons à une cadence donnée. Toutes les X heures, le tronçon de l'avion passe à la station suivante.

2. SOI pour *Standard Operating Instruction*

Zone : cette notion est relative à la définition du produit final. Les zones sont des espaces prédéfinis dans l'avion.

A cela vient s'ajouter un ensemble de contraintes temporelles. Pour le séquencement des gammes, chaque gamme est renseignée par des informations de précédences, correspondant à des contraintes techniques de nature multiples. Pour une gamme g_1, la précédence correspond à la liste des gammes qui doivent absolument être terminées avant de commencer g_1.

A ce jour le séquencement de l'ensemble des gammes, induit par la précédence, est décrit dans un diagramme de PERT[3] construit par un Agent d'Evaluation du Temps. Ce travail est réalisé manuellement, sans automatisation.

Dans notre cas d'étude, l'ordre de grandeur de la volumétrie traitée est de 500 gammes qui correspondent à 3500 opérations de montage de 10000 pièces (hors quincaillerie).

Pour commencer notre étude, nous avons compilé les différentes sources d'informations (base de données d'ingénierie, de production, agents de production et architectes industriels, documentation) à notre disposition et avons retenu :

– un plan de production sous forme de fichier MS Project, contenant le diagramme de PERT des gammes ;

– les gammes sous forme de document pdf, contenant non seulement les opérations mais aussi les zones d'interventions sur la section en cours d'assemblage, la liste des pièces à installer et le nombre d'opérateurs requis pour l'assemblage d'une section d'avion ;

– la liste, sous forme de fichiers de tableurs, des pièces à installer sur les différentes IP de la ligne pulsée (obtenue à partir de requêtes du système d'information de production).

L'étape suivante a consisté à analyser ces sources d'informations et, identifier les données pertinentes pour notre étude. Ensuite, nous avons élaboré un premier modèle de données établissant les liens entre ces données et réalisé une base de données pour les stocker. Pour finir nous avons analysé en détails et extrait de chaque gamme les données que nous avons saisies (manuellement) dans notre base de données (voir Figure 5).

Après analyses de toutes ces informations, nous pouvons conclure que nous avons deux mondes très différents avec, d'un côté, la conception, qui est un monde que nous pouvons qualifier de statique, très descriptif, avec des éléments, des positions et des tailles. De l'autre côté, la fabrication, qui est un monde de nature plus dynamique, constituée d'actions, d'opérations, comme visser un élément, et de durée.

3. PERT pour Program Evaluation and Review Technique, c'est un diagramme qui permet de déterminer le chemin critique qui conditionne la durée minimale d'une séquence d'actions.

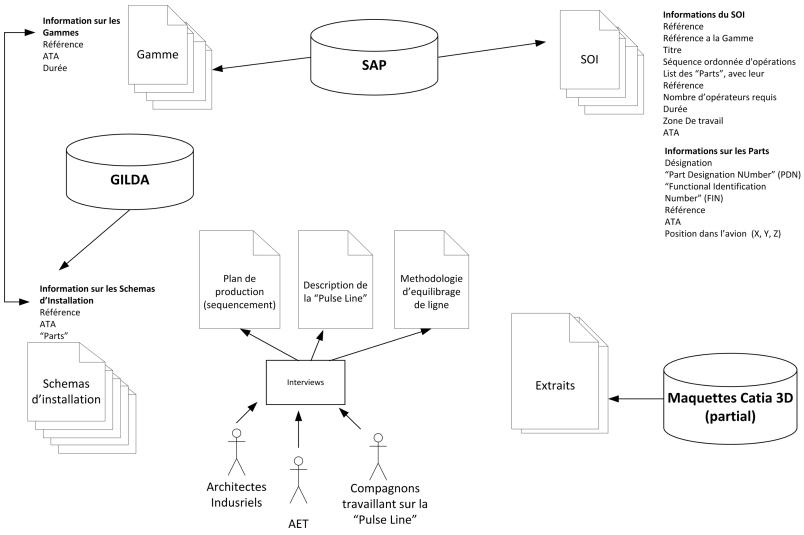


Figure 5. *Sources d'information*

4. Vers un modèle pour l'architecte avion

4.1. *Faire un lien entre un monde d'opération et un monde de description*

Afin de disposer d'un modèle permettant à l'architecte de réaliser une conception simultanée de l'avion et de sa production, il nous faut créer une correspondance entre des éléments dynamiques, les opérations de fabrication, et des éléments statiques. Prenons l'exemple simpliste d'une tuyauterie, dans le monde de la conception, elle correspond à une définition statique précise (géométrie, matériaux, etc.) et, dans le monde de la fabrication, aux opérations nécessaires a son installation. Le concept même de l'objet tuyauterie est finalement la passerelle entre ces deux mondes. Ainsi, les éléments physiques, constitutifs de l'avion, sont pour nous la zone de contact entre ces deux univers.

Nous pourrions réaliser cette jonction au moyen d'un couplage, comme des règles de transformation, entre deux modèles distincts (Herrmann *et al.*, 2007). Nous n'avons pas choisi cette solution et ce pour trois raisons. Premièrement, il existe un risque non négligeable d'une dérive entre les deux modèles. En effet, il est envisageable, qu'aux travers des évolutions futures dans la vie de l'entreprise, les modèles changent sans que leur couplage ne soit mis à jour, voire qu'ils deviennent complètement incompatibles entre eux. La deuxième raison est plus idéologique. Que cela soit en conception, comme en fabrication, il n'y a qu'un seul et unique avion : on ne fabrique pas un autre avion que celui conçu. Dès lors, il est préférable de montrer, au sein de l'entreprise, un modèle unique de l'avion qui peut être partagé par tous, plutôt que des modèles hétérogènes qui donnerait à penser que l'objet réel n'est pas exactement le même. Pour finir,

la troisième raison est liée à notre but. Toute notre démarche vise à permettre à l'architecte d'appréhender les interactions entre l'architecture et la production dans une vue cohérente et unique. Il ne s'agit donc pas d'avoir deux modèles distincts, conception et production, mais bien un modèle de l'avion qui dispose d'une partie conception et d'une partie moyen de production.

Cependant, nous ne voulons pas non plus fusionner de façon inextricable le modèle de production et celui de conception. Nous devons pouvoir envisager que, dans le futur, les modèles à chaque univers puissent évoluer, sans que chaque évolution de l'un impacte nécessairement l'autre. Notre idée est donc de définir un modèle, muni de deux parties, avec un nombre limité d'objets partagés entre les parties. Dans ce modèle, les objets partagés définissent la "*zone de contact*", la liaison, entre la partie production et la partie conception. Ainsi, chaque métier peut faire évoluer son modèle sans impacter l'autre du moment qu'il ne touche pas aux objets de liaisons.

Parallèlement, pour pouvoir travailler dans les premières phases du cycle de développement, quand les pièces ne sont pas encore connues ou définies en détails, il est nécessaire de travailler à un haut niveau d'abstraction. Aujourd'hui, cette vision existe déjà dans la conception. Avant de se décliner en pièces, l'avion est vu au travers de grands ensembles de composants. Ces justement ces différents niveaux d'abstraction, qui n'existent pas du côté fabrication, que nous cherchons à caractériser.

Dans la ligné de (Wisnosky et Vogel, 2004) avec les cadres d'architectures, nous avons choisi une approche en vues, ou chaque vue correspond à une couche d'abstraction. Par contre, contrairement aux approches macromodèles (Salay *et al.*, 2009) qui sont composées de couches d'abstraction avec un système explicitant le passage d'un niveau à un autre, nous avons pour le moment privilégié la simplicité, en considérant un modèle abstrait de haut niveau qui vient se spécialiser dans des classes concrètes pour chaque niveau d'abstraction.

4.2. *Notre modèle*

Une première ébauche de notre modèle conceptuel abstrait est présentée Figure 6. Ce modèle abstrait doit se spécialiser dans chaque vue, dans chaque niveau d'abstraction de la conception. Pour des raisons de lisibilité, nous avons choisi de faire figurer que les attributs et les classes les plus importantes. De plus, dans cet article, nous ne rentrerons pas dans les détails techniques de la conception et des liens avec les modèles 3D, ni dans les détails du moyen de production (source d'énergie, qualification et disponibilité de la main d'œuvre, etc.).

Nous avons fait figurer en haut de la figure les concepts relatifs à la conception et en bas ceux relatifs à la production. Le pont entre les deux mondes étant l'avion physique, représenté ici par la classe abstraite *Element physique*. Cette classe se spécialise aux travers des classes concrètes *Pièce* (voir Figure 7), pour un modèle de bas niveau, et en *Sous-élément* pour un modèle associée à la conception de haut niveau. La notion de Sous-élément est un concept que nous avons défini dans cette étude. Un

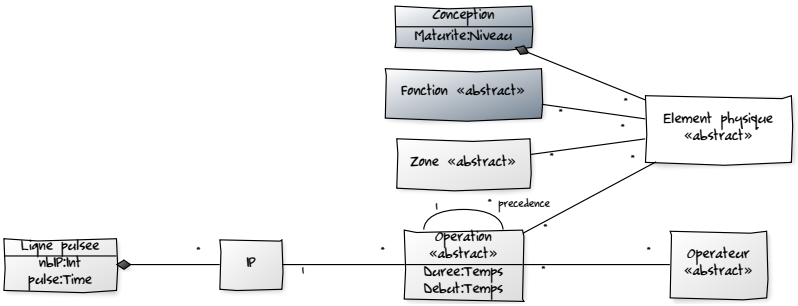


Figure 6. *Prototype de diagramme UML de classe générique pour un modèle de conception simultanée aéronautique*

Sous-élément est un ensemble de pièces qui a une pertinence propre à un stade du processus de production (résultat intermédiaire d'opérations d'assemblage). Un ensemble de sous-éléments définit un composant de l'avion qui correspond à une partie de l'avion (voilure, tronçon avant) fabriquée par différents sites de production et livré pour assemblage aux usines qui réalisent l'assemblage final.

Du côté conception, une proposition d'architecture correspond à un objet de type *Conception* auquel est associé un niveau de maturité et qui est composé d'éléments physiques. Chaque objet Elément Physique est associé à une *Fonction* avion. Ces fonctions correspondent à la vie opérationnelle de l'avion : générer de l'électricité, stocker le carburant, fournir les moyens de communication etc. Là encore, ces fonction vont se spécialiser :

– à haut niveau d'abstraction, par des fonctions qui correspondent aux services rendues aux utilisateurs de l'avion (pilotes, passagers, personnel cabine, etc.), sans présupposer de la façon dont ces fonctions sont réalisées,

– au niveau le plus détaillé, par des *Fonctions élémentaires* qui sont toutes les fonctions techniques nécessaires pour réaliser les fonctions avion et reflétant une solution technique.

Du côté production, nous avons deux classes concrètes qui représentent la chaîne de production : la *Ligne pulsée* et l'*IP*. La ligne pulsée a un temps de pulse, dit *TAKT*, qui correspond à la cadence, et un nombre d'IP. Sur chaque IP sont réalisées des Opérations. Une opération correspond à des actions que réalise un opérateur sur l'IP. Une opération a une durée et une date de début. A haut niveau d'abstraction, ces opérations vont se spécialiser sous la forme d'action à très gros grain qui consiste à équiper l'avion avec un Sous-élément sans rentrer dans les détails. Au niveau le plus fin, ces opérations correspondent à celle décrites en section 3 dites *Operation SOI*.

Pour finir, nous avons sur notre modèle générique une classe abstraite *Opérateur*, qui correspond à l'acteur qui réalise l'Opération, et *Zone* qui correspond à l'emplacement de l'Element physique.

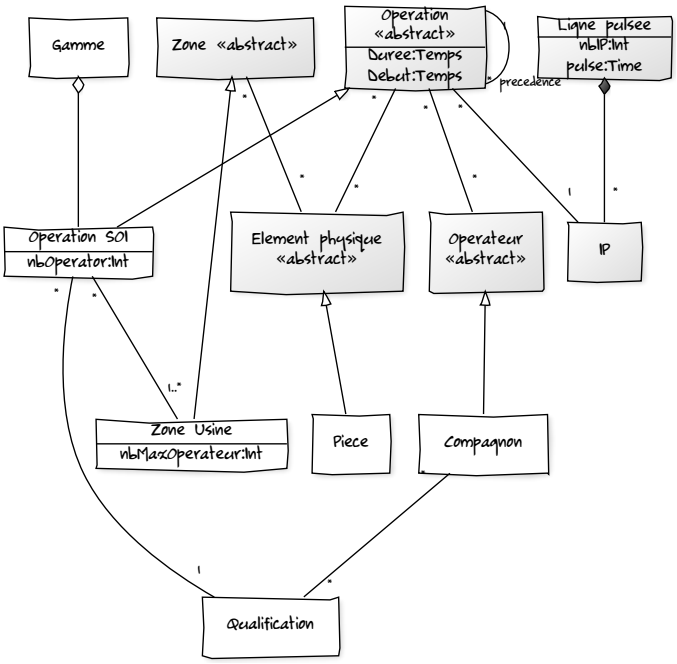


Figure 7. *Diagramme UML de classe spécialisé, partie fabrication*

Nous donnons Figure 7, un exemple de spécialisation de la partie production de notre modèle abstrait. Ici, le modèle actuel de conception est composé de descriptions fonctionnelles et physiques de l'avion. Chaque Pièce est décrite sous une forme 3D et est liée à des *Opérations SOI* et les informations concernant ces opérations sont décrites en détail dans les *Gammes*, documents destinés aux opérateurs devant réaliser les opérations de montage d'un ensemble de pièces au sein d'une zone sur une station. Les opérateurs qui réalisent les tâches sont des *Compagnon*, il n'y a pas de robot dans cet exemple, et chaque compagnon à des *Qualification*. La Zone est ici spécialisée en *Zone Usine* qui identifie l'endroit où les compagnons devront intervenir.

Si nous reprenons notre exemple du concept d'architecture de la conduite de conditionnement permettant de supprimer un goulot d'étranglement Section 2.2, nous avons les éléments suivants :

– la conduite de conditionnement d'air est de type Sous-élément et appartient à une vue de haut niveau ;

– les sections de tuyaux la constituant sont des objets de type Pièce. C'est au travers de ces objets que se fait le lien entre la conception (définition du tuyaux, forme, matériaux, etc.) et la production. Ils sont associés aux objets de type Opérations SOI, qui décrivent les opérations nécessaires pour l'installation de la pièce ;

– la pointe avant est une zone de haut niveau qui va se décomposer en objets de type Zone ;

– la modification de la route de la conduite de cheminement d'air va entraîner le déplacement des opérations nécessaires à son installation dans une autre zone de l'avion. Ces opérations peuvent être décrites de manière macroscopique lors des études d'architecture et de manière détaillée (classe Opération SOI) pour être fournies aux opérateurs de production.

5. Conclusion

Dans cet article, nous avons présenté un cas d'étude industriel, qui montre, pour la conception d'objets[4] complexes tel qu'un avion, l'importance de disposer d'un cadre de réflexion et de conception englobant aussi bien la dimension du produit comme celle de son système industriel. L'approche modèle permet de poser clairement les différents concepts, d'analyser le problème et d'avoir un socle commun pour faciliter la communication entre les équipes conception et fabrication. De plus, notre modèle est une première étape vers un double digital permettant, in fine, de réaliser des études d'impact d'architecture sur la production. Après une étude précise de la réalité de terrain de notre cas d'étude, qui a mobilisé l'essentiel de nos efforts jusqu'à présent, nous avons donné une première ébauche de ce qui sera notre modèle pour une conception simultanée de l'avion et de sa production.

Dans les travaux futurs, nous allons devoir affiner la partie production de notre modèle afin de pouvoir réaliser un simulateur de la chaîne production pouvant permettre à l'architecte d'estimer des cadences en fonction de choix de conception. Pour ce faire, nous pensons utiliser des outils issus de la recherche opérationnelle tels que (Pralet et Verfaillie, 2013). Par ailleurs, il pourrait être intéressant de faire un lien entre notre modèle statique, qui décrit le produit et le système de production, avec le processus de fabrication qui relève d'une vue dynamique. Pour cela, nous pourrons voir du côté de travaux qui cherchent à unifier ces deux vues aux travers de diagrammes SysML (Batarseh et McGinnis, 2012). Dans un tout autre registre, dans la lignée des travaux de (Bruno *et al.*, 2015), nous pourrions chercher à utiliser une ontologie orientée fabrication afin de permettre aux architectes d'avoir des raisonnements de haut niveau d'abstraction sur des données très détaillées, en établissant le lien entre la vision éléments physiques et les fonctions réalisées.

6. Bibliographie

Batarseh O., McGinnis L. F., « SysML to Discrete-event Simulation to Analyze Electronic Assembly Systems », *Proceedings of the 2012 Symposium on Theory of Modeling and Simulation - DEVS Integrative M&S Symposium*, TMS/DEVS '12, San Diego, CA, USA, 2012, Society for Computer Simulation International, p. 48 :1–48 :8.

4. Nous employons ici le terme d'objet plutôt que système, car nous voulons mettre en avant le caractère usinable, de système qui donne lieu à une fabrication industrielle.

Benkamoun N., ElMaraghy W., Huyet A.-L., Kouiss K., « Architecture Framework for Manufacturing System Design », *Procedia CIRP*, vol. 17, 2014, p. 88 - 93.

Bruno G., Antonelli D., Villa A., « A Reference Ontology to Support Product Lifecycle Management », *Procedia CIRP*, vol. 33, 2015, p. 41–46.

Decreuse C., Feschotte D., « Ingénierie simultanée », *Techniques de l'ingénieur Stratégies de conception pour l'innovation*, vol. base documentaire : TIB127DUO., 2017, Editions T.I.

Delmas R., Doose D., Pires A. F., Polacsek T., « Supporting Model Based Design », Bellatreche L., Pinto F. M., Eds., *Model and Data Engineering - First International Conference, MEDI 2011, Óbidos, Portugal, September 28-30, 2011. Proceedings*, vol. 6918 de *Lecture Notes in Computer Science*, Springer, 2011, p. 237–248.

Delmas R., Polacsek T., « Formal Methods for Exchange Policy Specification », Salinesi C., Norrie M. C., Pastor O., Eds., *Advanced Information Systems Engineering - 25th International Conference, CAiSE 2013, Valencia, Spain, June 17-21, 2013. Proceedings*, vol. 7908 de *Lecture Notes in Computer Science*, Springer, 2013, p. 288–303.

Demoly F., Yan X., Eynard B., Rivest L., Gomes S., « An assembly oriented design framework for product structure engineering and assembly sequence planning », *Robotics and Computer Integrated Manufacturing*, vol. 27, n° 1, 2011, p. 33–46.

Göpfert I., Schulz M., « *Logistics Integrated Product Development in the German Automotive Industry : Current State, Trends and Challenges* », p. 509–519, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

Graham R. L., « Bounds for certain multiprocessing anomalies », *The Bell System Technical Journal*, vol. 45, n° 9, 1966, p. 1563-1581.

Herrmann C., Krahn H., Rumpe B., Schindler M., Völkel S., « An Algebraic View on the Semantics of Model Composition », Akehurst D. H., Vogel R., Paige R. F., Eds., *Model Driven Architecture- Foundations and Applications, Third European Conference, ECMDA-FA 2007, Haifa, Israel, June 11-15, 2007, Proccedings*, vol. 4530 de *Lecture Notes in Computer Science*, Springer, 2007, p. 99–113.

Pralet C., Verfaillie G., « Dynamic Online Planning and Scheduling Using a Static Invariant-Based Evaluation Model », Borrajo D., Kambhampati S., Oddi A., Fratini S., Eds., *Proceedings of the Twenty-Third International Conference on Automated Planning and Scheduling, ICAPS 2013, Rome, Italy, June 10-14, 2013*, AAAI, 2013.

Salay R., Mylopoulos J., Easterbrook S. M., « Using Macromodels to Manage Collections of Related Models », van Eck P., Gordijn J., Wieringa R., Eds., *Advanced Information Systems Engineering, 21st International Conference, CAiSE 2009, Amsterdam, The Netherlands, June 8-12, 2009. Proceedings*, vol. 5565 de *Lecture Notes in Computer Science*, Springer, 2009, p. 141–155.

Shenas D. G., Derakhshan S., « Organizational approaches to the implementation of simultaneous engineering », *International Journal of Operations & Production Management*, vol. 14, n° 10, 1994, p. 30–43, MCB UP Ltd.

Sprock T., McGinnis L. F., « Analysis of Functional Architectures for Discrete Event Logistics Systems (DELS) », *Procedia Computer Science*, vol. 44, 2015, p. 517 - 526.

Wisnosky D. E., Vogel J., « DoDAF Wizdom : A Practical Guide to Planning », *Managing and Executing Projects to Build Enterprise Architectures Using the Department of Defense Architecture Framework (DoDAF)*, , 2004.

Alignement, union et intersection de modèles : 3 transformations pour l'analyse des systèmes d'information

André Miralles[1], **Marianne Huchard**[2], **Jessie Carbonnel**[2], **Clémentine Nebut**[2]

1. *Tetis/IRSTEA, France*
   andre.miralles@teledetection.fr
2. *LIRMM, CNRS & Université de Montpellier, France*
   marianne.huchard,jessie.carbonnel,clementine.nebut@lirmm.fr

RÉSUMÉ. *En système d'information, l'intégration de modèles consiste à regrouper au sein d'un unique modèle l'ensemble des entités métiers de plusieurs modèles connectés d'un point de vue thématique. Dans cette communication, trois transformations sont proposées afin d'assister cette intégration et d'améliorer les modèles : la première produit un modèle d'alignement montrant les correspondances entre les modèles, la seconde produit un modèle union (Least Common Multiple Model ou LCM) et la troisième produit un modèle intersection (Greatest Common Model ou GCM) constitué des seuls éléments communs à tous les modèles. Le LCM est la plus petite des unions des modèles et le GCM, noyau de tous les modèles, est la plus grande des intersections des modèles. Ces transformations sont réalisées à l'aide de l'Analyse Formelle de Concepts (AFC) en basant les transformations sur des opérations portant sur les contextes formels.*

ABSTRACT. *In information systems, model integration consists in grouping into a single model all the business entities of several thematically connected models. In this paper, three transformations are proposed to assist this integration: an alignment model which highlights the correspondences between the models; a union model (Least Common Multiple Model or LCM) and the intersection model (Greatest Common Model or GCM) built up with the common elements. The LCM is the smallest union of the models and the GCM, kernel of all models, is the greatest intersection of the models. These transformations are achieved with the help of Formal Concept Analysis (FCA).*

MOTS-CLÉS : *Système d'information, UML, modèle de classes, appariement de modèles, intégration de modèles, union de modèles, intersection de modèles, Analyse Formelle de Concepts*

KEYWORDS: *Information System, UML, Class model, class model matching, class model integration, class model union, class model intersection, Formal Concept Analysis*

1. Introduction

Quelle que soit l'application informatique à développer et a fortiori le système d'information, l'analyse du système et des besoins des utilisateurs est probablement la phase la plus délicate car d'elle dépend le succès ou l'échec d'un développement informatique. Au cours de cette phase, la capture des *entités métiers*[1] par le concepteur reste la tâche la plus sensible car ce dernier doit s'approprier les entités pour le capturer afin de les retranscrire et de les décrire correctement dans le modèle de l'application.

L'ingénierie dirigée par les modèles a pour but d'automatiser via des transformations l'évolution des modèles depuis l'analyse jusqu'à la génération de code. Dans ces conditions, le modèle est alors le principal produit de l'analyse. Une conséquence directe est que le modèle doit être parfaitement structuré et avoir une qualité sémantique irréprochable. Dans la pratique, sauf pour des modèles de faible taille, il est impossible de satisfaire ces exigences de qualité en quelques séances d'analyse. Cela a conduit à la multiplication des méthodes de développement itératives ou agiles (Kruchten, 1999; Beck, 2000; Alliance, 2001) qui permettent d'améliorer progressivement la structuration ou la qualité sémantique.

La réalisation de ce modèle "idéal" est d'autant plus difficile que le nombre d'acteurs impliqués est grand et que les domaines (thématiques, scientifiques, législatifs, etc.) concernés sont nombreux et variés. Dans ce contexte, (Miralles, 2016) préconise de faire l'analyse par petits groupes homogènes d'acteurs. Un autre avantage de cette analyse en groupe est que, si certains acteurs sont en situation de conflit potentiel (ex. agriculteur/écologiste), l'analyse sera plus sereine et donc de meilleure qualité. La contrepartie de l'analyse en groupe est la nécessité de disposer de méthodes et d'outils pour fusionner les modèles produits en un seul. Cette problématique d'intégration se rencontre dans d'autres situations (Miralles, 2016) et en particulier lors de l'inventaire de l'existant. S'il existe déjà des applications ou des bases de données dans des domaines connexes, les ateliers de génie logiciel actuels permettent de reconstruire un modèle UML par ré-ingénierie du code ou de la base de données. Le concepteur dispose alors de plusieurs modèles UML à intégrer. Plusieurs auteurs ont développé des méthodes et des outils pour réaliser cette factorisation soit par des techniques d'alignement de schémas (Batini *et al.*, 1986; Rahm, Bernstein, 2001; Shvaiko, Euzenat, 2005) soit, plus récemment, en mettant en œuvre l'Analyse Formelle de Concepts (Stumme, Maedche, 2001; Amar *et al.*, 2012). Dans cet article, nous proposons une approche mettant en œuvre l'Analyse Formelle de Concepts pour mener à bien cette intégration en effectuant différentes transformations. Par rapport à un alignement de schémas qui produit habituellement seulement des correspondances entre éléments de modélisation, nous construisons trois types de modèles (schémas) : le modèle d'alignement qui est le plus proche de l'alignement de schémas standard, le LCM (Least Common Multiple Model) qui est l'intégration minimale et complète

1. Afin d'éviter toute confusion entre la notion de `Concept métier` utilisée en modélisation et celle de `Concept` d'un treillis, le terme `Concept métier` est remplacé par `Entité métier`.

des modèles et le GCM (Greatest Common Model) qui est la factorisation maximale des modèles. Ces deux derniers modèles bornent au sens mathématique l'espace de correspondance des modèles sources. Ces trois modèles peuvent être construits indépendamment.

Le contexte et la problématique sont présentés en section 1. La section 2 rappelle le principe et les produits résultant de l'Analyse Formelle de Concepts qui permettent de reconstruire des modèles d'alignement, d'union et d'intersection obtenus à partir des contextes formels associés, contextes qui sont définis en section 3. La section 4 illustre l'approche par un cas d'étude. Enfin, avant d'aborder la conclusion en section 6, la section 5 présente les travaux connexes qui ont inspiré et alimenté notre réflexion.

2. Eléments sur l'Analyse Formelle de Concepts

Notre approche prend sa source dans l'analyse par treillis (Barbut, Monjardet, 1970), également connue comme Analyse Formelle de Concepts (Ganter, Wille, 1999). L'AFC est une approche mixte permettant à la fois (1) l'extraction de connaissances et de règles dans des données et leur ordonnancement par spécialisation/généralisation, (2) la construction de classifications conceptuelles telles que des ontologies, des modèles entités-relations et des modèles de classes. Ces propriétés lui sont procurées par sa capacité à mettre en évidence des schémas communs et des différences dans les données soumises à l'analyse.

Dans sa forme la plus simple, un ensemble ordonné de concepts est formé à partir d'un *contexte formel* composé d'un ensemble d'entités décrites par des caractéristiques. Un *contexte formel* est ainsi un triplet $K = (G, M, I)$, où G est l'ensemble d'entités, M l'ensemble de caractéristiques et $I \subseteq G \times M$ une relation binaire dont chaque couple (g, m) associe une entité g à une caractéristique m qu'elle possède. La Table 1 présente un contexte formel associant aux classes du modèle UML de la Figure 2 (modèle M1[2]), leurs attributs et les rôles qu'elles possèdent dans des associations. Dans cette table, par exemple, la classe `AcidRain` est associée aux attributs `timePoint`, `codeQuality`, `waterAmount` et `particuleAmount`. Les trois premiers attributs sont hérités, tandis que le quatrième est déclaré dans la classe elle-même. La classe `RainEvent` est associée quant à elle à l'attribut `timePeriod` et au rôle `storedRain`.

Pour un contexte formel $K = (G, M, I)$ donné, un *concept formel* $C = (Extent(C), Intent(C))$ associe un ensemble maximal d'entités avec un ensemble maximal de caractéristiques qu'elles possèdent. L'extension du concept est l'ensemble $Extent(C) = \{g \in G | \forall m \in Intent(C), (g, m) \in I\}$ des entités couvertes par le concept, tandis que son intension $Intent(C) = \{m \in M | \forall g \in Extent(C), (g, m) \in I\}$ contient les caractéristiques partagées. Par exemple le concept $Concept\_M1\_6$=({-

2. Dans ce qui suit, M1 et M2 désignent respectivement les modèles `M1 Weather Station Model` et `M2 Weather Station Model`.

Table 1. Contexte formel K_{M1}

M1	**serialNumber**	**tubeHeight**	**timePoint**	**codeQuality**	**waterAmount**	**timePeriod**	**acquisitionFrequency**	**accuracyClass**	**windStrength**	**windDirection**	**particuleAmount**	**measuredRain**	**storedRain**	**measuredWind**
RainGauge	×	×										×		
Rain			×	×	×									
RainEvent						×							×	
Anemometer							×	×						×
Wind			×	×					×	×				
AcidRain			×	×	×						×			

$Wind, Rain, AcidRain\}, \{timePoint, codeQuality\}$) regroupe trois classes partageant les deux attributs de son intension (les exemples sont extraits de la Figure 1).

Étant donnés deux concepts formels $C_1 = (E_1, I_1)$ et $C_2 = (E_2, I_2)$ d'un contexte formel K, un ordre de spécialisation/généralisation $\preceq_C$ peut être donné par $C_1 \preceq_C C_2$ si et seulement si $E_2 \subseteq E_1$ (et de manière équivalente $I_1 \subseteq I_2$). C_1 est une spécialisation (un sous-concept) de C_2. C_2 est une généralisation (un super-concept) de C_1. Par exemple $Concept\_M1\_5 = (\{Rain, AcidRain\}, \{timePoint, codeQuality, waterAmount\})$ est un sous-concept de $Concept\_M1\_6$.

Par ces définitions, C_1 hérite des caractéristiques de C_2, tandis qu'inversement, C_2 hérite des entités de C_1. Les caractéristiques et entités héritées sont souvent omises sur les schémas pour des raisons de lisibilité. C'est ainsi que la représentation graphique de $Concept\_M1\_6$ ne présente que les caractéristiques introduites $\{timePoint, codeQuality\}$ et celle de $Concept\_M1\_5$ ne présente que la caractéristique introduite $\{waterAmount\}$ et l'entité introduite $\{Rain\}$. Une caractéristique (resp. une entité) est dite *introduite* par un concept s'il s'agit du concept le plus général (resp. le plus spécifique) où elle apparaît. Si $\mathcal{C}_K$ est l'ensemble de tous les concepts de K, le munir de la relation d'ordre $\preceq_C$ lui confère une structure de treillis.

Plusieurs applications de l'AFC considèrent uniquement le sous-ordre du treillis restreint aux concepts qui introduisent au moins une caractéristique ou au moins une entité. Ce sous-ordre porte le nom d'AOC-poset et se trouve souvent utilisé dans les applications relatives à la classification conceptuelle. La Figure 1 présente l'AOC-poset associé au contexte formel de la Table 1. L'interprétation d'un AOC-poset dans un tel contexte sera reprise plus en détail dans la section 4. Pour en donner un premier aperçu ici, on peut constater que les concepts offrent une vue par spécialisation des classes. Ces dernières apparaissent dans les extensions des concepts (partie basse des boîtes), tandis que les attributs ou rôles introduits apparaissent dans les intensions des concepts (partie centrale des boîtes). La construction de l'AOC-poset met en évidence, grâce au $Concept\_M1\_6$, une possible super-classe de `Wind` et `Rain`, factorisant `timePoint` et `codeQuality`, et introduisant le concept métier de *mesure*. Elle montre aussi (ce qui était déjà connu) que `AcidRain`, introduite dans un sous-concept de celui qui introduit `Rain`, doit en être une sous-classe. Quelquefois ces

relations ne sont pas explicites dans le modèle et elles le deviennent dans l'AOC-poset. Dans la section suivante, nous étudions une nouvelle utilisation de l'AFC, cette fois en présence de plusieurs modèles UML, pour analyser leurs correspondances et leur possible intégration.

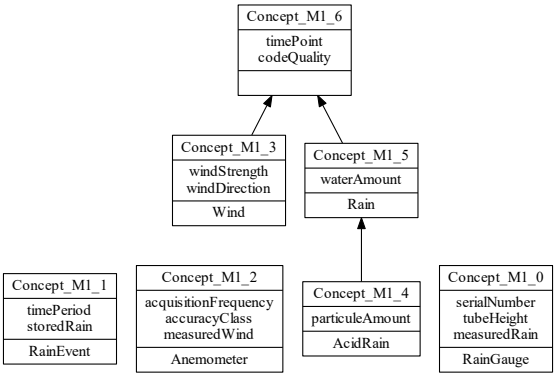


Figure 1. AOC-poset du contexte formel K_{M1} de station météorologique

3. Contextes formels pour l'alignement, l'union et l'intersection de modèles

Trois contextes formels vont nous permettre d'analyser respectivement l'alignement, l'union et l'intersection de modèles. Différentes hypothèses de construction de ces contextes peuvent être faites. Ici, nous supposons que la terminologie des modèles a été uniformisée au préalable. Ainsi deux classes (resp. deux attributs) portant le même nom seront supposées représenter les mêmes notions et avoir une même sémantique.

Le *contexte formel d'alignement*, présenté en haut de la Table 2, consiste à concaténer les contextes formels associés aux modèles M1 et M2, après avoir préfixé les noms des classes du nom du modèle dont elles proviennent. Dans cet article, pour des raisons de place nous ne présentons pas séparément le contexte formel du modèle M2. Il peut être retrouvé facilement à partir de la partie inférieure de la Table 2 (partie haute) : il correspond à toutes les lignes préfixées par "M2-". Une ligne de ce contexte correspond donc à une paire (modèle d'origine, classe). Une colonne correspond à un attribut ou un rôle provenant de l'un ou de l'autre des contextes (ou des deux).

Dans le *contexte formel union*, présenté au centre de la Table 2, on retrouve les colonnes du contexte formel d'alignement. Par contre, les lignes du contexte d'alignement qui correspondent à des classes portant des noms identiques sont fusionnées. Dans notre exemple, ce sera le cas pour `M1-RainGauge` et `M2-RainGauge`, fusionnées dans la ligne `M1::M2-RainGauge` ou pour `M1-Anemometer` et `M2-Anemometer`, fusionnées dans la ligne `M1::M2-Anemometer`. Pour ces classes "com-

Table 2. Contextes formels pour l'alignement, l'union et l'intersection sémantique

ALIGNEMENT	**serialNumber**	**tubeHeight**	**timePoint**	**codeQuality**	**waterAmount**	**timePeriod**	**acquisitionFrequency**	**accuracyClass**	**windStrength**	**windDirection**	**particuleAmount**	**measuredRain**	**storedRain**	**measuredWind**	**anemometerType**	**measuredSnowFall**
M1-RainGauge	×	×										×				
M1-Rain			×	×	×											
M1-RainEvent						×							×			
M1-Anemometer							×	×						×		
M1-Wind			×	×					×	×						
M1-AcidRain			×	×	×						×					
M2-RainGauge		×										×				
M2-Precipitation			×	×	×											
M2-SnowFall				×	×											
M2-Anemometer							×	×						×	×	
M2-Breeze			×						×							
M2-SnowGauge		×														×

LCM (Union)	**serialNumber**	**tubeHeight**	**timePoint**	**codeQuality**	**waterAmount**	**timePeriod**	**acquisitionFrequency**	**accuracyClass**	**windStrength**	**windDirection**	**particuleAmount**	**measuredRain**	**storedRain**	**measuredWind**	**anemometerType**	**measuredSnowFall**
M1::M2-RainGauge	×	×										×				
M1-Rain			×	×	×											
M1-RainEvent						×							×			
M1::M2-Anemometer							×	×						×	×	
M1-Wind			×	×					×	×						
M1-AcidRain			×	×	×						×					
M2-Precipitation			×	×	×											
M2-SnowFall				×	×											
M2-Breeze			×						×							
M2-SnowGauge		×														×

GCM (Intersection)	**tubeHeight**	**acquisitionFrequency**	**accuracyClass**	**measuredRain**	**measuredWind**
M1::M2-RainGauge	×			×	
M1::M2-Anemometer		×	×		×

munes", la description est l'union des descriptions (attributs et rôles) des deux modèles. Par exemple, bien que `M1-Anemometer` ne possède pas `anemometerType`, comme `M2-Anemometer` le possède, cet attribut est associé à `M1::M2-Anemometer` dans l'union.

Dans le *contexte formel intersection*, présenté au bas de la Table 2, on ne garde que les classes communes et les colonnes du contexte formel d'alignement qui sont associées à une classe commune aux deux modèles et ceci dans les deux modèles. Ainsi, l'attribut `serialNumber`, qui n'est associé à `RainGauge` que dans le modèle M1,

n'apparaît pas. Une classe n'est donc associée à un attribut ou à un rôle que si elle le possède dans les deux modèles.

4. Cas d'étude

Dans le monde agricole, les prévisions météorologiques sont des informations majeures pour les agriculteurs car d'elles dépendent la croissance des plantes, mais aussi certains travaux agricoles à effectuer et, en particulier, les traitements phytosanitaires. Ces prévisions issues de mesures locales sont de plus en plus souvent « personnalisées » afin de bien répondre aux besoins des agriculteurs. Face à la multiplication de produits proposés, le souhait des acteurs agricoles est de disposer d'un modèle unique issu des différents modèles de station météorologique existants. L'étude cas portera sur la fusion de deux d'entre eux.

4.1. Présentation des deux modèles à fusionner

La station météorologique, dont le modèle est celui de la Figure 2, est équipée d'un pluviomètre (`RainGauge`) pour effectuer des relevés ponctuels de pluie (`Rain`) et d'un anémomètre (`Anemometer`) pour mesurer les caractéristiques du vent (`Wind`). Le pluviomètre est identifié par un numéro de série (`serialNumber`) et la hauteur du tube (`tubeHeight`) recueillant l'eau de pluie, paramètre propre à ce matériel. L'anémomètre est quant à lui caractérisé par sa fréquence d'acquisition (`acquisitionFrequency`) ainsi que par sa classe de précision (`accuracyClass`). Les quantités de précipitation (`waterAmount`) ainsi que l'intensité et la direction du vent (`windStrength` et `windDirection` respectivement) sont datées (`timePoint`) et renseignées par un code de qualité globale de la mesure (`codeQuality`). Cette station permet en outre de suivre la quantité de particules (`particuleAmount`) des pluies acides (`AcidRain`) qui sont un cas particulier de pluie. Pour reconstituer un événement pluvieux (`RainEvent`), il est nécessaire d'identifier, pendant la durée de l'épisode (`timePeriod`), les différents relevés ponctuels de pluie effectués, via le rôle `storedRain`.

Le modèle de la Figure 3 est celui d'une station météorologique permettant en outre de relever les chutes de neige (`SnowFall`) au moyen d'un nivomètre (`SnowGauge`). Ce dernier instrument est un appareil qui donne l'équivalent en eau de la quantité de neige recueillie. De ce fait, comme pour le pluviomètre, la hauteur du tube (`tubeHeight`) recueillant la neige est une caractéristique intrinsèque de l'appareil. La similarité se retrouve aussi dans les caractéristiques relevées qui sont la quantité d'eau (`waterAmount`) et un code global de qualité (`codeQuality`) de cette mesure. Pour cette station, ces mesures ne sont pas datées (absence de `timePoint`) car, comme les chutes de neige sont rares, le choix des gestionnaires a été de reporter la date sur un cahier de terrain. Par ailleurs, la connaissance des événements pluvieux n'est pas nécessaire pour les utilisations envisagées (absence d'une entité `RainEvent`). En outre, cette station n'est pas équipée pour le suivi des pluies

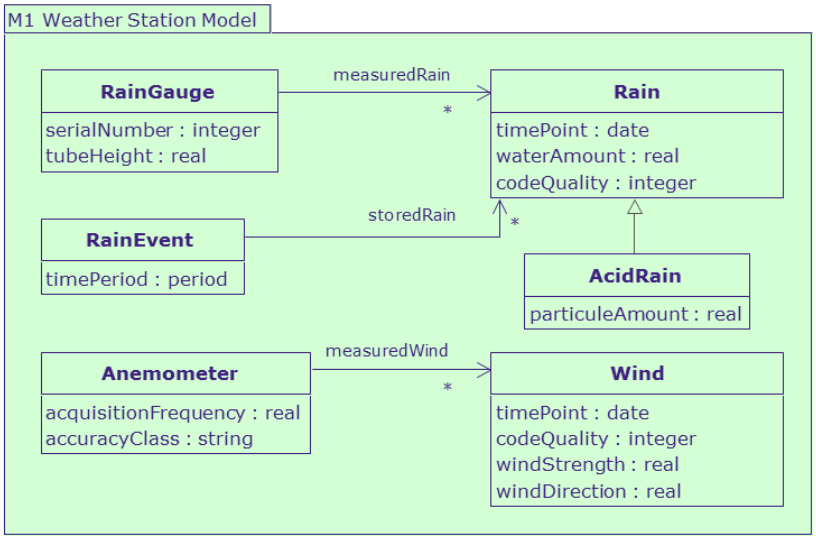


Figure 2. Modèle de la station météorologique M1

acides (modèle sans la classe `AcidRain`). Les autres différences importantes par rapport au modèle de la station M1 sont :

– l'utilisation d'entités sémantiquement différentes pour désigner la pluie et le vent ; dans le modèle M2, l'entité pluie (`Rain`) est remplacée par précipitation (`Precipitation`) et vent (`Wind`) par brise (`Breeze`),

– le numéro de série du pluviomètre (`serialNumber`) n'est pas pris en compte,

– la direction du vent (`windDirection`) est ignorée ainsi que la qualité de la mesure (`codeQuality`),

– enfin, une information importante pour le suivi à long terme de cette station est le type d'anémomètre utilisé, type qui peut évoluer dans le temps et expliquer certains écarts de mesure.

4.2. Alignement des modèles M1 et M2

L'AOC-poset du contexte formel d'alignement (Table 2) des modèles M1 et M2 est donné en Figure 4. Cet AOC-poset permet de reconstruire le modèle UML de la Figure 5 représentant l'alignement des modèles M1 et M2.

Une première analyse rapide de l'AOC-poset permet de constater la présence de trois nouveaux concepts ($Concept\_Alignment\_8$, $Concept\_Alignment\_11$ et $Concept\_Alignment\_12$) issus du calcul de l'AFC. Ces trois nouveaux concepts sont le résultat de la factorisation des attributs `tubeHeight`, `timePoint` et `codeQuality` respectivement. Ces concepts doivent être validés et nommés par un expert du domaine et, dans le cas présent, ils donnent naissance à trois nouvelles entités

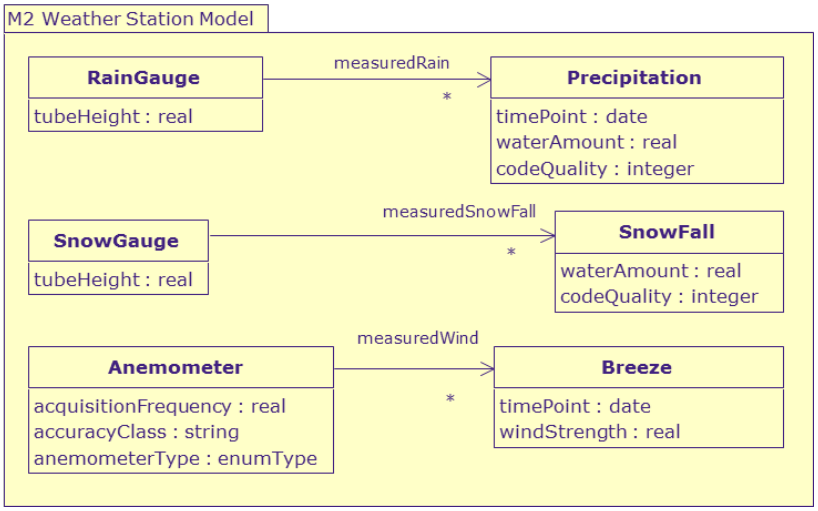


Figure 3. Modèle de la station météorologique M2

métiers qui sont `PrecipitationDevice`, `TimeParameter` et `QualityParameter` respectivement (cf. Figure 5).

L'AFC fusionne aussi les entités métiers `Rain` du modèle M1 et `Precipitation` du modèle M2 au sein du concept $Concept\_Alignment\_9$. Ce n'est pas surprenant au vu des attributs les décrivant. Ce concept représente l'entité métier `Rain/Precipitation` de la Figure 5. Les quatre entités métiers évoquées ci-dessus sont extérieures aux modèles M1 et M2 car elles sont nouvelles.

Dans l'AOC-poset de la Figure 4, les entités métiers `Anemometer` de M1 et M2 apparaissent dans deux concepts distincts ($Concept\_Alignment\_3$ et $Concept\_Alignment\_7$) donnant ainsi naissance à deux classes dans le modèle d'alignement de la Figure 5. L'entité métier `Anemometer` de M2 spécialise celle de M1 car elle possède en plus l'attribut `anemometerType`. Il en est de même pour les deux entités métiers `RainGauge` de M1 et de M2 qui partagent le rôle `measuredRain` ($Concept\_Alignment\_6$) alors que l'attribut `serialNumber` est propre à l'entité métier `RainGauge` de M1 ($Concept\_Alignment\_0$), d'où la relation de spécialisation en Figure 5. La présence des doubles occurrences de `Anemometer` et de `RainGauge` impose le maintien des modèles M1 et M2 au sein du modèle d'alignement (`Alignment Model`) de la Figure 5.

L'entité métier `Anemometer` est en relation avec `Wind` au travers d'une association dont le rôle est `measuredWind`. Le concept $Concept\_Alignment\_5$ factorise l'attribut `windStrength` qui est commun aux entités métiers `Wind` et `Breeze` de M1 et M2 respectivement. L'entité `Wind` spécialise `Breeze` puisque `Wind` a en plus l'attribut `windDirection` (cf. Figure 5).

Enfin, les entités métiers `RainEvent` avec son attribut `timePeriod` et `AcidRain` qui a comme attribut `particuleAmount` sont propres au modèle M1. `RainEvent` a une association vers l'entité fusionnée `Rain/Precipitation` qui a pour rôle `storedRain`. `AcidRain` spécialise l'entité fusionnée `Rain/Precipitation` comme c'était le cas dans le modèle M1. Les entités métiers `SnowGauge` et `SnowFall` appartiennent au modèle M2. Elles sont reliées par une association dont le rôle de `SnowFall` est `measuredSnowFall`.

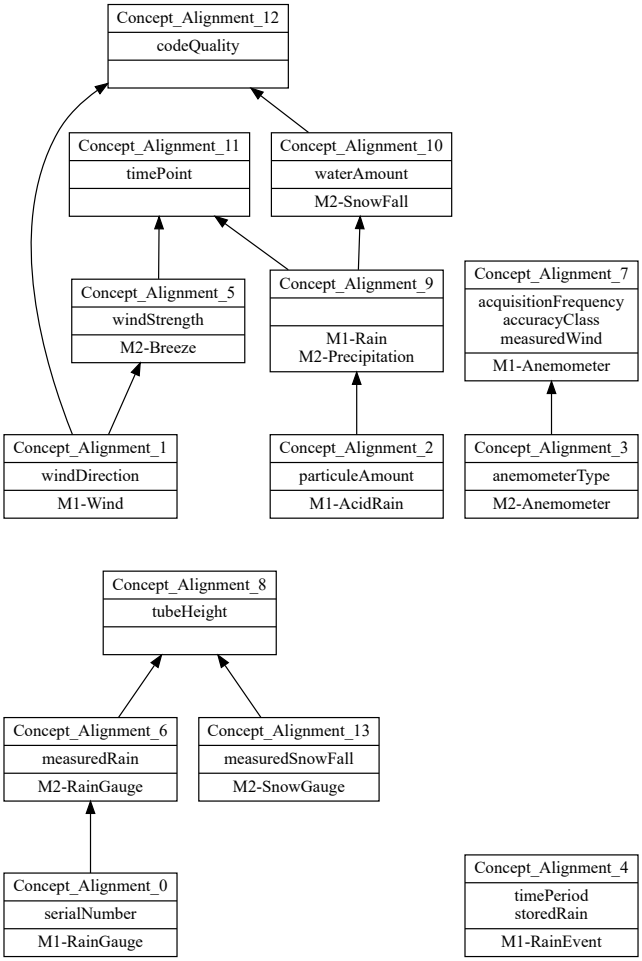


Figure 4. AOC-poset du contexte formel d'alignement

4.3. Union des modèles M1 et M2

La Figure 6 montre l'AOC-poset du contexte formel d'union (Table 2) des modèles M1 et M2. Par construction (cf. section 3), le contexte formel d'union est quasi-

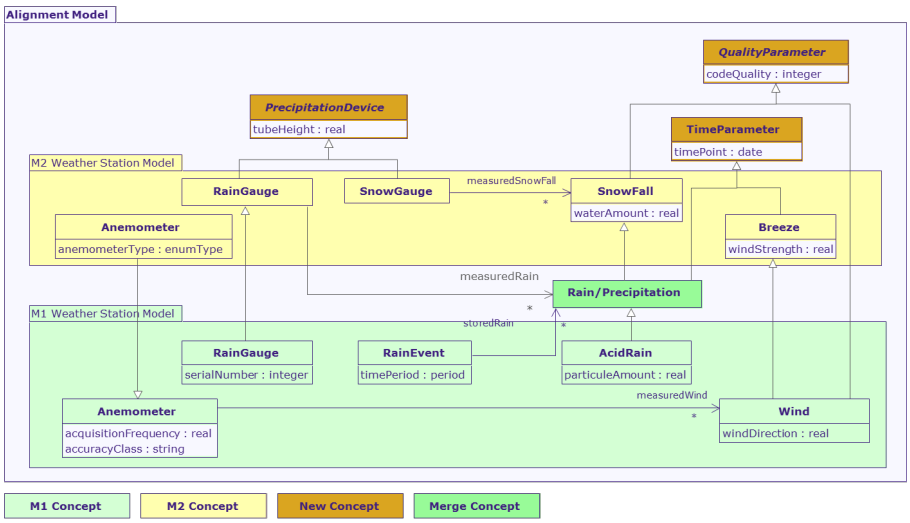


Figure 5. Modèle d'Alignement des modèles M1 et M2

ment identique à celui du contexte formel d'alignement (cf. Figure 5). Au nom et à la numérotation des concepts près (identifiants générés automatiquement), les seules différences résident dans :

– la fusion des concepts $Concept\_Alignment\_3$ et $Concept\_Alignment\_7$ de la Figure 5 pour donner le concept $Concept\_Union\_4$ de la Figure 6, concept où on retrouve les propriétés des entités métiers `Anemometer` des deux modèles M1 et M2,

– l'assimilation des deux concepts $Concept\_Alignment\_0$ et $Concept\_Alignment\_6$ de la Figure 5 au sein du concept $Concept\_Union\_0$ de la Figure 6 ; les propriétés des entités `RainGauge` des deux modèles M1 et M2 sont réunies dans le concept $Concept\_Union\_0$.

Il en résulte que le modèle LCM de la Figure 7, issu du calcul du contexte formel d'union, est plus simple et plus facile à analyser que le modèle d'alignement de la Figure 5. Cela se traduit par un gain de productivité lors de l'analyse approfondie puisque celle-ci est plus rapide. Une analyse approfondie comme celle du modèle d'alignement (cf. 4.2) aboutirait plus rapidement au même résultat comme ce sera présenté en section 4.5.

4.4. Intersection des modèles M1 et M2

L'AOC-poset du contexte formel d'intersection (Table 2) des modèles M1 et M2 est représenté en Figure 8. Il est composé des deux concepts $Concept\_Intersection\_0$ et $Concept\_Intersection\_1$. Le premier de ces concepts regroupe les deux entités métiers `RainGauge` des modèles M1 et M2 ainsi que les seules propriétés communes à ces deux entités : l'attribut `tubeHeight` et le rôle `measuredRain`.

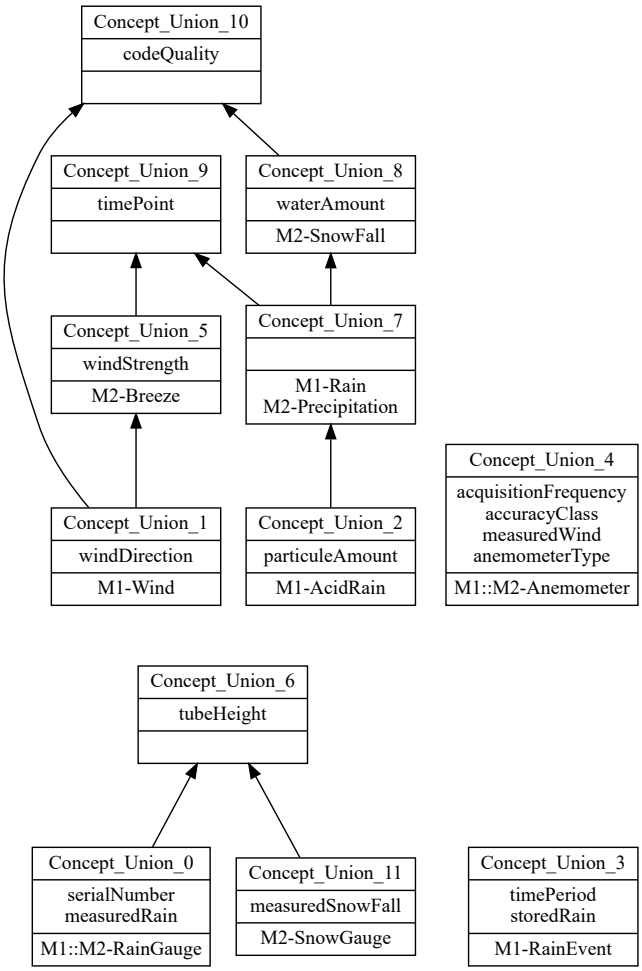


Figure 6. AOC-poset du contexte formel d'union

Par définition, l'attribut `serialNumber` n'est pas un élément de cette intersection, car il n'apparaît pas dans le modèle M2.

Le second concept correspond à la fusion des entités `Anemometer` des deux modèles. Comme pour l'entité `RainGauge`, les seules propriétés communes sont les attributs `acquisitionFrequency` et `accuracyClass` et le rôle `measuredWind`. `AnemometerType` étant une propriété de la seule entité `Anemometer` de M2, elle n'est pas un élément du GCM.

La Figure 9 montre le GCM reconstruit à partir de l'AOC-poset de la Figure 8. Dans le GCM, la classe `RainGauge` n'a que l'attribut `tubeHeight` et une association vers une classe à définir et à nommer mais dont le rôle est `measuredRain`. De façon similaire, la classe `Anemometer` possède les deux attributs `acquisitionFre-`

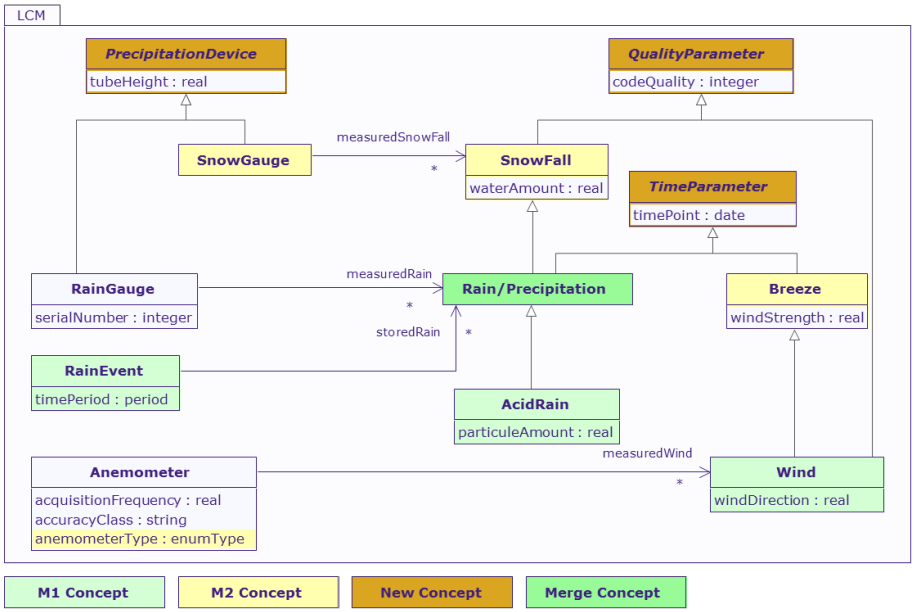


Figure 7. LCM des modèles M1 et M2

`quency` et `accuracyClass` ainsi qu'une association vers une classe non définie ayant pour rôle `measuredWind`.

Même si, pour des raisons de réutilisation du code, l'algorithme développé calcule le modèle LCM à partir du modèle d'alignement et le modèle GCM à partir du modèle LCM, ces trois modèles peuvent être construits indépendamment.

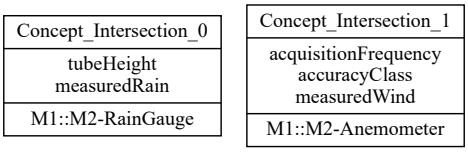


Figure 8. AOC-poset du contexte formel d'intersection

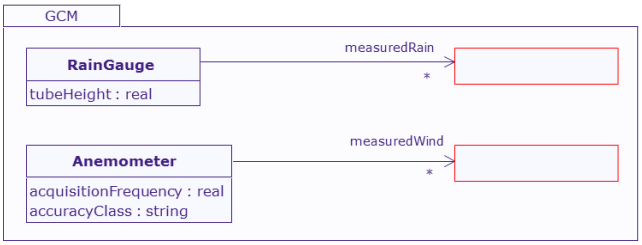


Figure 9. GCM des modèles M1 et M2

4.5. Ré-analyse pour l'intégration des modèles M1 et M2

Une fois les étapes de reconstruction des modèles d'alignement, d'union et d'intersection terminées, il est possible de poursuivre l'analyse des modèles pour les améliorer et surtout pour aboutir à un modèle intégré, plus générique et plus cohérent. Le modèle de la Figure 10 est le résultat de cette analyse approfondie, analyse qui ne peut se faire qu'avec le concours d'un expert du domaine. En premier lieu, l'analyse du modèle d'alignement remet en cause le choix des gestionnaires d'inscrire la date des relevés de neige sur un cahier de terrain. Cette pratique n'étant pas fiable à long terme, la décision est prise d'ajouter une date (`timePoint`) à l'entité métier `SnowFall`. Il est alors possible de fusionner les classes `QualityParameter` et `TimeParameter` et de les remplacer par l'entité `Data` qui a comme attributs `timePoint` et `codeQuality`. La poursuite de l'analyse met en évidence que les entités `Anemometer` peuvent être regroupées en une seule classe. La même opération peut être réalisée pour les entités `RainGauge`. N'ayant plus de classe en doublon, il est alors possible de supprimer les paquetages M1 et M2. Comme, du point de vue sémantique, les entités métiers `Rain` et `SnowFall` sont des cas particuliers de l'entité `Precipitation`, il est normal de les fusionner en une seule et même entité `Precipitation`. L'entité métier `AcidRain` reste quant à elle inchangée et spécialise l'entité fusionnée `Precipitation`. Afin que le modèle d'alignement réanalysé soit sémantiquement cohérent, d'une part, la classe `RainEvent` est renommée `PrecipitationEvent`, entité de niveau d'abstraction plus élevé, et d'autre part, le rôle de l'association entre les classes `PrecipitationEvent` et `Precipitation` devient `storedPrecipitation`. Le numéro de série (`serialNumber`) n'est pas une propriété propre à `RainGauge`. Elle est généralisable à tout instrument de mesure. A ce titre, elle s'applique à `Anemometer` et à `SnowGauge`. Cela conduit à créer une nouvelle classe `Device` contenant l'attribut `serialNumber` généralisant les classes `Anemometer` et `PrecipitationDevice`. Les associations ayant pour rôle `measuredRain` et `measuredSnowFall` sont généralisées par une association entre les classes `PrecipitationDevice` et `Precipitation`. Le rôle de cette nouvelle association est remplacé par `measuredPrecipitation`, abstraction des précédents rôles. Enfin, les entités métiers `Breeze` et `Wind` étant sémantiquement semblables, il est possible de les fusionner en une seule entité `Wind` ayant un niveau d'abstraction plus élevé que `Breeze`.

5. Travaux connexes

Les treillis et l'AFC ont été utilisés en génie logiciel et en base de données pour différents objectifs et, en particulier, pour construire, maintenir ou réorganiser des hiérarchies de classes ou des schémas de bases de données par refactorisation (Missikoff, Scholl, 1989; Rundensteiner, 1992; Godin, Mili, 1993; Snelting, Tip, 2000; Huchard, 2015). Certains autres travaux s'intéressent aussi au typage et à la multiplicité des propriétés et à la navigation des associations (Roume, 2004; J.-R. Falleri, 2009; Osman-Guédi, 2013).

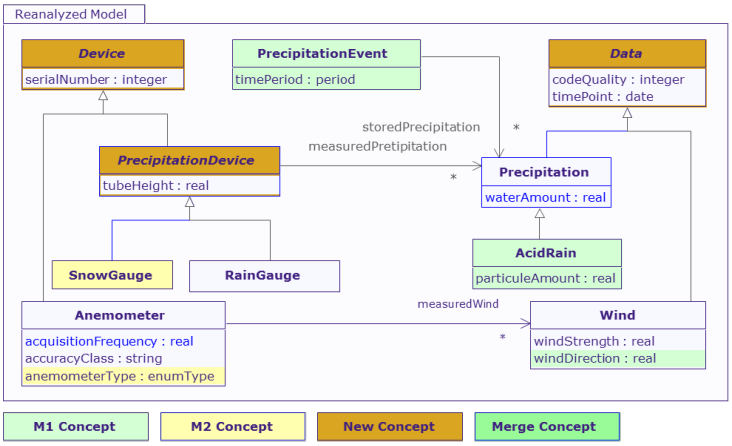


Figure 10. Modèle d'alignement réanalysé

La présente approche est différente puisqu'ici l'AFC est une analyse inter-modèles par opposition aux approches intra-modèle précédemment développées. Dans l'approche inter-modèles de cet article, nous conservons la traçabilité des entités en préfixant leur nom par l'identifiant du modèle d'origine. Une approche consistant à construire le modèle GCM a été présentée dans (Amar *et al.*, 2012). Elle organise également les concepts métiers construits ou modifiés pour leur choix par un utilisateur. Elle a été testée sur des modèles de systèmes d'information portant sur les Pesticides. Ici, nous proposons, en plus du modèle GCM, une vue sur l'alignement et l'union de modèles.

Les problèmes d'intégration ou d'alignement de bases de données ont été largement étudiés, notamment pour produire un schéma global (Batini *et al.*, 1986; Rahm, Bernstein, 2001; Shvaiko, Euzenat, 2005). Toutefois, les travaux d'interopérabilité des données entre systèmes multi-bases de données posent des problèmes de même nature (Parent, Spaccapietra, 1998). Pour ces deux types de problématiques, l'intégration des bases et l'interopérabilité des données consistent à trouver des correspondances entre concepts des différents schémas et à les ré-organiser dans une même structure. Dans notre approche, nous nous focalisons au niveau modèle/schéma alors que (Parent, Spaccapietra, 1998) concentrent leur réflexion au niveau des données. Cette approche d'alignement met en valeur les correspondances, tandis que le GCM montre les sous-systèmes strictement identiques et que le LCM sert de support à une possible ré-organisation.

L'identification de similarités entre modèles ou entre méta-modèles a été étudiée dans le domaine de la gestion de versions (Altmanninger *et al.*, 2009), du développement distribué (Cicchetti *et al.*, 2008), ou pour assister le développement de transformation de modèles (J. Falleri *et al.*, 2008; Voigt, Heinze, 2010).

Le problème présenté ici se rapproche également de l'appariement ou de l'alignement d'ontologies, que certaines approches traitent avec l'AFC (Kalfoglou, Schorlemmer, 2005; Bendaoud *et al.*, 2008). L'approche présentée dans (Stumme, Maed-

che, 2001) utilise l'AFC et une analyse linguistique pour fusionner des ontologies dans le contexte du Web sémantique. Dans (Formica, 2006), l'alignement utilise une mesure de similarité basée sur l'AFC, tandis que dans (Tatsiopoulos, Boutsinas, 2009), l'alignement s'appuie sur la structure interne de l'ontologie et l'extraction de règles d'association. Toutes ces approches se focalisent sur la recherche de correspondances, elles peuvent servir pour affiner la construction des contextes formels d'alignement, de LCM ou de GCM. A partir des contextes formels, nous allons au-delà de la simple mise en évidence de correspondances, et nous proposons de nouvelles vues sur les concepts métiers et de nouveaux concepts plus abstraits. Tous les concepts métiers sont intégrés dans une structure de spécialisation.

6. Conclusion

L'alignement est une technique utilisée depuis de nombreuses années pour factoriser, assembler ou fusionner plusieurs schémas de bases de données ou des modèles en un seul. Son automatisation via l'AFC produit des modèles d'alignement qui sont complexes car, en particulier, les entités métiers de même nom dans différents modèles ne sont pas fusionnées directement. Le modèle LCM qui réunit tous les éléments des modèles sources élimine cet inconvénient et rend plus facile la ré-analyse du modèle. C'est la plus petite des unions des modèles sources. Le modèle GCM est quant à lui constitué des seuls éléments communs aux modèles sources. C'est la plus grande des intersections des modèles sources, c'est-à-dire le noyau de tous les modèles. Il est incontournable puisque tous les acteurs, codes, schémas de bases, etc. utilisent les éléments du GCM. Ce noyau est le capital de base pour le développement d'une nouvelle application ou d'un nouveau système d'information.

Les modèles d'alignement, LCM et GCM sont des modèles obtenus automatiquement (par les algorithmes développés dans le cadre de ce travail). Pour rendre utilisables ces modèles (obtenir le modèle ré-analysé), l'expertise humaine est indispensable comme le montre la section 4.2 pour le nommage des nouvelles entités mais aussi la section 4.5 où la fusion ou la réorganisation de certaines entités ne peut être faite que par un expert du domaine.

Une particularité de l'AFC dans ce contexte est que cette méthode produit simultanément une factorisation intra-modèle et une factorisation inter-modèles qu'il est difficile de découpler. Dans ce contexte-là, la seule solution à notre connaissance est l'intervention d'un expert pour faire des réarrangements intra ou inter-modèles. Toutefois, il est aussi possible d'assister l'expert à l'aide d'un arbre de décision comme présenté dans (Amar *et al.*, 2012).

Les travaux vont se poursuivre dans plusieurs directions. Tout d'abord, on peut définir des opérations d'union et d'intersection basées sur les n-uplets d'éléments (attributs et rôles) et non sur les noms des classes, ce qui peut faire émerger des informations complémentaires. De plus, la capacité de factorisation de l'AFC étant limitée, nous comptons mettre en œuvre l'Analyse Relationnelle de Concepts (ARC) qui étend

l'analyse aux relations. Enfin, lorsque plus de deux modèles doivent être intégrés, nous envisageons de définir un mode itératif d'intégration.

Dans cette étude de cas, il n'y a pas d'ambiguïté sur le nom des éléments de modélisation (classes, attributs, etc.) car le modèle est de petite taille et relève d'un domaine unique. Pour des modèles plus conséquents impliquant plusieurs domaines, il est fort probable de rencontrer les ambiguïtés (par exemple la Forêt d'un domaine forestier et le Foret utilisé pour faire des perçages). Dans ce cas, nous envisageons de conditionner la factorisation à une analyse des éléments de modélisation environnants mais aussi d'utiliser des ontologies métiers, des ressources lexicales et des outils de Traitements Automatiques du Langage. Dans (Carbonnel *et al.*, 2017), nous utilisons une variante de l'approche présentée ici en construisant les modèles d'alignement, LCM et GMC non plus sur la sémantique des noms des éléments de modélisation mais en s'appuyant sur les caractéristiques définissant les entités.

Nous allons poursuivre l'étude sur des modèles déjà mobilisés dans des travaux antérieurs (Amar *et al.*, 2012; Miralles *et al.*, 2014) pour évaluer la pertinence des alignements en termes de précision et de rappel.

References

Alliance A. (2001). *Manifesto for agile software development* (Vol. 2005) No. June.

Altmanninger K., Seidl M., Wimmer M. (2009). A survey on model versioning approaches. *International Journal of Web Information Systems*, Vol. 5, No. 3, pp. 271–304.

Amar B., Guédi A. O., Miralles A., Huchard M., Libourel T., Nebut C. (2012). Finding Semi-Automatically a Greatest Common Model Thanks to Formal Concept Analysis. In *Revised selected papers of ICEIS 2012, LNBIP vol. 141*, pp. 72–91.

Barbut M., Monjardet B. (1970). *Ordre et classification (volume 2)*. Hachette.

Batini C., Lenzerini M., Navathe S. B. (1986). A comparative analysis of methodologies for database schema integration. *ACM Computer Survey*, Vol. 18, pp. 323–364.

Beck K. (2000). *extreme programming explained - embrace change*. Addison-Wesley.

Bendaoud R., Napoli A., Toussaint Y. (2008). Formal Concept Analysis: A unified framework for building and refining ontologies. In *EKAW 2008*, p. 156-171.

Carbonnel J., Huchard M., Miralles A., Nebut C. (2017). Feature Model composition assisted by Formal Concept Analysis. In *To appear in 12th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE) 2017, April 28 - 29, 2017, Porto, Portugal.*

Cicchetti A., Ruscio D., Pierantonio A. (2008). Managing model conflicts in distributed development. In *MoDELS 2008*, pp. 311–325.

Falleri J., Huchard M., Lafourcade M., Nebut C. (2008). Metamodel Matching for Automatic Model Transformation Generation. In *MoDELS 2008*, pp. 326–340.

Falleri J.-R. (2009). *Contributions à l'IDM : reconstruction et alignement de modèles de classes*. Thèse de doctorat, Université Montpellier 2.

Formica A. (2006). Ontology-based concept similarity in Formal Concept Analysis. *Information Sciences*, Vol. 176, pp. 2624–2641.

Ganter B., Wille R. (1999). *Formal concept analysis: Mathematical foundation*. Springer-Verlag Berlin.

Godin R., Mili H. (1993). Building and maintaining analysis-level class hierarchies using Galois lattices. In *OOPSLA*, pp. 394–410.

Huchard M. (2015). Analyzing inheritance hierarchies through formal concept analysis: A 22-years walk in a landscape of conceptual structures. In *MASPEGHI@ECOOP 2015, ACM Digital Library*, pp. 8–13.

Kalfoglou Y., Schorlemmer M. (2005). Ontology mapping: The state of the art. In *Semantic interoperability and integration, Dagstuhl Seminar proceedings.*

Kruchten P. B. (1999). *The rational unified process: An introduction* (3rd ed.). Addison-Wesley.

Miralles A. (2016). *Contribution à une conception rationnelle et malléable des systèmes d'information environnementaux.* Habilitation à diriger les recherches. (Français)

Miralles A., Dolques X., Huchard M., Le Ber F., Libourel T., et. al. (2014). Exploration de la factorisation d'un modèle de classes sous contrôle des acteurs. In *Actes du XXXIIème Congrès INFORSID, Lyon, France, 20-23 Mai 2014*, pp. 245–261.

Missikoff M., Scholl M. (1989). An Algorithm for Insertion into a Lattice: Application to Type Classification. In *Proceedings of the 3rd Int. Conf. FODD 1989*, pp. 64–82.

Osman-Guédi A. (2013). *Transformation automatisée de modèles de Systèmes d'Information*. Thèse de doctorat, Université Montpellier 2.

Parent C., Spaccapietra S. (1998, May). Issues and approaches of database integration. *Communication of the ACM*, Vol. 41, pp. 166–178.

Rahm E., Bernstein P. A. (2001). A survey of approaches to automatic schema matching. *The VLDB Journal*, Vol. 10, pp. 334–350.

Roume C. (2004). *Analyse et restructuration de hiérarchies de classes*. Thèse de doctorat, Université Montpellier 2.

Rundensteiner E. A. (1992). *A Class Classification Algorithm For Supporting Consistent Object Views*. Technical report. University of Michigan.

Shvaiko P., Euzenat J. (2005). A Survey of Schema-Based Matching Approaches Journal on Data Semantics IV. In *Journal on data semantics IV*, Vol. 3730, pp. 146–171.

Snelting G., Tip F. (2000). Understanding class hierarchies using concept analysis. *ACM Trans. Program. Lang. Syst.*, Vol. 22, No. 3, pp. 540–582. Retrieved from http://doi.acm.org/10.1145/353926.353940

Stumme G., Maedche A. (2001). Ontology merging for federated ontologies on the semantic web. In *Int. work. foundations of models for information integration (FMII)*, pp. 413–418.

Tatsiopoulos C., Boutsinas B. (2009). Ontology mapping based on association rule mining. In *ICEIS 2009*, p. 33-40.

Voigt K., Heinze T. (2010). Metamodel matching based on planar graph edit distance. In *ICMT 2010*, pp. 245–259.

Filtrage d'informations

Filtrage collaboratif sensible au contexte

Une approche basée sur LDA

Josiane Mothe[1] , Ambinintsoa Jocelyn Rakotonirina[2]

1.ESPE, Université de Toulouse, Université de Toulouse Jean Jaurès
Institut de Recherche en Informatique de Toulouse, IRIT
UMR 5505 CNRS
118 Route de Narbonne, Toulouse, France
Josiane.Mothe@irit.fr

2. DMI, Université d'Antananarivo, Madagascar
Mathématiques Informatique et Statistique Appliquées, MISA
BP 906 Ankatso
Ambinintsoa26@outlook.com

RESUME. Les systèmes de recommandations visent à proposer aux utilisateurs des items en lien avec leur consultation en cours et qui peuvent retenir leur intérêt. L'intérêt des utilisateurs dépend du contexte dans lequel ils se trouvent. Dans ce travail, nous proposons un système hybride CBCF (Contex-aware Based Collaborative Filtering) qui combine le système de recommandations sensibles aux contextes et le filtrage collaboratif. Le contexte est ici défini comme l'objectif ou l'intention de l'utilisateur. Nous le modélisons par une approche LDA (Latent Dirichlet Allocation) qui génère un modèle de thèmes pour chaque intention. Nous avons évalué notre approche sur la collection Book-Crossing et montrons sa supériorité par rapport à plusieurs méthodes état de l'art.

MOTS-CLES : Système d'information, Recherche d'information, Accès à l'information, Système de recommandation, Latent Dirichlet Allocation, Filtrage collaboratif, Système de recommandation hybride

ABSTRACT. Recommender systems are designed to provide user with items related to their ongoing browsing and that may be of interest to them. User interest depends on the context. In this work, we propose a hybrid CBCF (Context-aware Based Collaborative Filtering) system combining context-sensitive and collaborative filtering. We define context as the objective or intent of the user. We model it by a LDA (Latent Dirichlet Allocation) approach which generates a topic model for each intention. We evaluated our approach using the Book-Crossing collection and demonstrated the superiority of our model over several state-of-the-art methods.

KEYWORDS: Information Systems, Information Retrieval, Recommender systems, Latent Dirichlet Allocation, Collaborative filtering, Hybrid recommender system.

1. Introduction

Depuis le début des années 1990, internet a changé la manière de consommer et de vendre : l'e-commerce est devenu un moyen commun de commerce. Sur un site de e-commerce, l'enjeu pour les entreprises est d'attirer plus de clients, de les aider à accéder rapidement aux items (produits, services, films, restaurants, etc.) pertinents et de transformer une visite sur le site en un achat.

Les systèmes de recommandations (SR) sont une solution pour recommander automatiquement des items aux utilisateurs qui peuvent être perdus dans un vaste choix. Les systèmes développés pour répondre à cet enjeu améliorent l'expérience client et augmentent le chiffre d'affaire des e-commerces (30% du chiffre d'affaire en 2011 chez Amazon.com selon Nick Tsinonis au sein de RecSys.com).

De nombreux travaux se sont intéressés aux SR mais certains défis restent à lever encore aujourd'hui comme le démarrage à froid qui désigne un manque d'information lors de l'ajout d'un nouvel utilisateur ou d'un nouvel item au système [Schein et al., 2002], la rareté ou la parcimonie des données explicites comme les notes des utilisateurs qui souvent n'évaluent pas les items [Adomavicius et Tuzhilin, 2005] et le manque, voire l'absence de diversité dans les recommandations des items [Chevalier *et al.*, 2016][Candillier *et al.*, 2011]. Au risque d'être intrusif, un SR se doit aussi d'être le plus pertinent possible pour le client. Le système devrait s'adapter aux situations car souvent les données sur les entités (utilisateurs, produits, etc.) sont dynamiques et évoluent [Louëdec *et al.*, 2015].

Dans la littérature les SR sensibles aux contextes sont utilisés pour traiter ce caractère variable des préférences. Selon [Dey, 2001], un contexte désigne n'importe quelle information qui peut caractériser la situation d'une entité (personne, produit, localisation, etc.). [Palmisano *et al.*, 2008] ont analysé l'influence des informations contextuelles dans la prédiction des comportements et dans la modélisation des utilisateurs (l'étude définit les contextes comme le but ou l' intention d'achats des utilisateurs dans un SR). En fait, les auteurs ont étudié le comportement des utilisateurs qui est susceptible de changer dans différents contextes. En effet, pour un site e-commerce, différents clients peuvent acheter un même produit pour différentes intentions. Par exemple, un champagne peut être considéré comme un produit de luxe adapté pour un cadeau par exemple, mais pour d'autres consommateurs il s'agit d'un produit essentiel pour une fête. Si le champagne est vu par les utilisateurs comme boisson de luxe, ils trouveront pertinents la recommandation d'autres produits de luxe, mais s'il est vu comme produit de fêtes, d'autres accessoires de fêtes seront pertinents. La notion de contexte est étduiée par ailleurs dans de nombreux domaines comme la prise en compte du contexte métier dans l'accès à l'information [Chaker et al., 2013] ou la prise en compte du contexte dynamique dans les profils utilisateurs [Canut et al., 2015].

Les études dans la plupart des SR utilise les notes des utilisateurs pour trouver la similarité entre les items et ne considère pas le contexte dans lequel se trouve l'utilisateur au moment de noter. Dans cet article, nous proposons un SR sensible au

contexte utilisant les descriptions des items pour trouver la similarité entre ces items. Pour atteindre cet objectif, nous combinons deux approches :

- la modélisation des thèmes, qui permet de rechercher l'intention des utilisateurs à partir des descriptions textuelles d'un ou plusieurs items successifs qu'ils ont consultés ;
- un système de filtrage collaboratif basé sur le thème de l'utilisateur (ou profil utilisateur) extrait précédemment.

La suite de cet article est structurée comme suit. La section 2 présente l'état de l'art. La section 3introduit la motivation de la méthode que nous proposons, les jeux de données choisis, l'implémentation et l'évaluation de l'approche. La section 4 montre les résultats empiriques et les analyses. La section 5 conclut cet article.

2. Etat de l'art

Les SR peuvent être définis comme des programmes qui visent à recommander les éléments ou items à partir de l'item qui est consulté par l'utilisateur et des informations connexes. Trois types de SR ont été définis dans la littérature [Adomavicius *et al.*, 2005]. Les SR basés sur le contenu génèrent les recommandations à partir de l'historique des préférences de l'utilisateur associé aux caractéristiques (description, prix, couleur, etc.) des items courants [Pazzani et Billsus, 2007]. Le filtrage collaboratif forme un groupe d'utilisateurs qui a les mêmes préférences, ainsi, seuls les items les plus appréciés par le groupe sont pertinents [Adomavicius *et al.*, 2005]. Enfin, l'approche hybride combine les deux approches précédentes [Burke, 2007].

Selon [Adomavicius et Tuzhilin, 2011] malgré un nombre considérable de recherches faites sur les SR, la plupart des approches se focalisent sur la recommandation des items les plus pertinents pour les utilisateurs sans prendre en compte les informations contextuelles (exemple : le temps, la localisation ou la compagnie d'autres personnes). [Adomavicius et Tuzhilin, 2011] ont montré que les informations contextuelles pertinentes ont des influences importantes sur un SR. Il est donc important d'étudier les SR sensibles aux contextes.

La notion de contexte est intéressante pour les SR. Selon [Dey, 2001], un contexte est n'importe quelle information qui peut caractériser la situation d'une entité (personne, localisation, produit, etc.). [Ryan *et al.*, 1999] définissent le contexte comme l'identité de l'utilisateur, ressources de l'environnement proche, localisation de l'utilisateur et période temporelle d'exécution de l'interaction. Selon [Berry et Linoof, 1997], les contextes sont définis comme des évènements qui caractérisent les phases de la vie d'un client et qui peuvent influencer ses préférences, son statut et sa valeur pour une entreprise. Des études comportementales en marketing ont montré que la prise de décision des clients dépend des contextes dans lesquels ils se trouvent [Adomavicius et al., 2005]. En effet, selon les contextes comme la localisation, les saisons, l'humeur, etc. le même client peut choisir différents produits.

Plusieurs recherches ont été menées dans différents domaines pour évaluer l'impact des contextes dans les SR. Ces recherches ont été faites sur les contextes observables (localisation, compagnie, période, etc.) [Borras *et al.*, 2014][Lamsfus *et al.*, 2009] et les contextes non observables (identité d'un membre d'une famille)[Palmisano *et al.*, 2008].

[Adomavicius et al., 2005] ont présenté un SR avec une méthode multidimensionnelle. Des contextes sont ajoutés à la fonction d'évaluation de dimension deux (R : User x Item → Rating). Ainsi on obtient une fonction multidimensionnelle (R : User x Item x Contexte → Rating) qui inclut les informations contextuelles dans la prédiction des préférences des utilisateurs. Pour implémenter la méthode multidimensionnelle et tester sa performance, des données sur des films (notes) et des données contextuelles (localisation, période, compagnie) ont été collectées. Ces données contextuelles ne sont pas disponibles sur la collection de référence Movielens [movielens.umn.edu] généralement utilisé pour évaluer les SR, ni sur les autres données publiques. Par conséquent, un site internet spécifique a été créé et il a été demandé à des utilisateurs d'évaluer les films qu'ils ont vus ainsi que les informations contextuelles pertinentes. Les résultats montrent empiriquement une amélioration de la prédiction des films des systèmes sensibles aux contextes par rapport aux systèmes qui ne les incluent pas.

Selon [Borras et al., 2014] les activités des voyageurs touristiques peuvent être variables en temps réel ; il faut donc adapter les recommandations aux circonstances des voyages (exemple : il pleut ou pas, à l'intérieur ou à l'extérieur d'un musée). Ainsi, dans les applications qui utilisent la mobilité (tourisme, visite de musée, restauration, etc.), les SR sensibles aux contextes améliorent l'expérience utilisateur. L'approche développée par [Lamsfus et al., 2009] utilise les contextes (localisation, période, météo courant) et propose des suggestions à tout instant en fonction des préférences d'activités du touriste. Par exemple, si un client s'attarde sur une activité qu'il rencontre sur la route et que le temps pour les autres activités a du retard, alors les visites suivantes devraient être adaptées au plan initial.

Les informations contextuelles proviennent de plusieurs sources diversifiées. Ainsi, caractériser un contexte de recommandations d'items se différencie par rapport à ses origines. Selon [Adomavicius et Tuzhilin , 2011] il y a trois manières d'obtenir les informations contextuelles :

(1) Explicitement, en posant directement des questions aux utilisateurs (sondages) qui utilisent un site web.

(2) Implicitement, par les informations sur les achats effectués, le nombre de clics, la localisation de l'utilisateur grâce aux smartphones (utilisé en tourisme, restauration).

(3) Par induction, en utilisant des modèles prédictifs (ou des classsifieurs). Par exemple dans un supermarché, il est difficile de connaître explicitement l'identité d'un membre d'une famille, qui réalise des achats ensemble avec une seule carte de paiement ou un même compte pour l'e-commerce. Avec les méthodes d'induction utilisant les classifieurs Naïves Bayes et les

réseaux bayésiens, [Palmisano et al., 2008] ont montré que, des informations contextuelles cachées (ici identité d'un membre) peuvent être induites à partir des données existantes (ici les items achetés).

Un défi se pose sur cette dernière façon d'obtenir les informations contextuelles. Le problème qui se pose est qu'il est difficile de modéliser les contextes à partir des informations contextuelles non observables comme l'intention de l'utilisateur. Une alternative pour le résoudre est la modélisation de thèmes que nous proposons de réaliser via une modélisation LDA.

3. Filtrage collaboratif basé sur LDA

Modéliser les contextes à partir d'informations contextuelles non observables comme l'intention d'achat de l'utilisateur est difficile. Nous avons choisi de capturer l'intention implicite de l'utilisateur par les thèmes associés aux items qu'il a consultés. La méthode LDA (Latent Dirichlet Allocation) permet cette modélisation des thèmes.

Selon [Blei, 2012] les modèles de thèmes sont des techniques d'apprentissage automatique et statistiques qui analysent les mots des textes dans les documents pour découvrir les thèmes qu'ils traitent, comment ces thèmes sont connectés entre eux et comment ils changent au fil du temps.

LDA a été appliqué dans l'analyse de documents [Griffiths et Steyvers, 2004] [Fei-Fei et Perona, 2005], la catégorisation et le regroupement de documents [Wei et Croft ; 2006] [Ramage *et al.*, 2009] et l'ordonnancement de systèmes de recherche d'information [Ionescu *et al.*, 2015]. LDA a été introduit dans les SR afin d'analyser le contexte dans les méthodes basées sur le contenu [Yu *et al.*, 2012]. Dans cet article, nous proposons une autre utilisation du modèle LDA. Le résultat du modèle LDA est intégré dans un système de filtrage collaboratif pour trouver la similarité entre les items consultés par un utilisateur courant.

Nous définissons le profil d'un utilisateur en le représentant par son thème. Notre approche recommande alors des items pour lesquels les distributions de thèmes des titres sont similaires au profil de l'utilisateur courant.

Ainsi, cet article propose un SR sensible aux contextes ; il s'agit d'un modèle de recommandation hybride qui combine la méthode de modélisation de thèmes basée sur LDA et la méthode de filtrage collaboratif. Notre approche est décomposée en trois étapes qui sont décrites dans les sous-sections suivantes.

3.1. Modèle LDA pour la représentation des thèmes des utilisateurs

La première étape implémente le modèle LDA à partir des descriptions des items que les utilisateurs ont consultés. LDA est utilisé pour extraire la structure sémantique cachée dans les descriptions des items que les utilisateurs ont consultés, la distribution des mots sur les thèmes latents et le mélange des distributions des thèmes latents. Cela consiste à estimer la distribution de thèmes latents (noté Θ)

pour chaque item et la distribution de mots (noté φ) pour chaque thème. Ces distributions vont permettre d'identifier la sémantique de l'espace de thèmes latents en les rapportant aux mots et aux items.

Dans la littérature, l'algorithme EM (Expected Maximization) [Blei *et al.*, 2003] et l'algorithme Gibbs sampling [Griffiths et Steyvers, 2004] sont les méthodes les plus utilisées pour l'estimation des paramètres (distributions) Θ et φ du modèle LDA. Cependant, l'algorithme EM est pénalisé par un grand nombre d'opérations à cause du grand nombre de documents ; il est donc plus lent à converger. L'algorithme Gibbs sampling permet de contourner cette difficulté. C'est pour cette raison que nous l'avons utilisé dans ce travail.

Le Collapsed Gibbs Sampling [Griffiths et Steyvers, 2004] est un algorithme d'échantillonnage qui permet l'estimation des paramètres d'un espace discret de grande dimension [Steyvers et al., 2004].

Dans cet article, la méthode de Gibbs sampling est utilisée pour estimer les paramètres de LDA qui itèrent plusieurs fois sur chaque mot v pour extraire un nouveau thème k pour le mot basé sur la probabilité p (z_i=k| v_i, z_{-i}) comme suit :

$$p(z_i=k|v_i,z_{-i}) \propto (n_{d,k}+\alpha_k)\frac{n_{k,v}+\beta_v}{\sum_{v'} n_{k,v'}+\beta_{v'}} \qquad (1)$$

Où $n_{k,v}$ calcule le nombre des affectations thème-mot.

$n_{d,k}$ calcule le nombre des affectations document-thème.

z_{-i} désigne toutes les affectations thème-mot et document-thème sauf pour l'affectation courantez_i pour le mot v_i.

α et β sont les paramètres de Dirichlet utilisés comme des paramètres de lissage pour les calculs.

A partir de l'équation (1). Les paramètres θ et φ du modèle LDA sont estimées comme suit [Griffiths et Steyvers, 2004] :

$$\theta_{d,k} = \frac{n_{d,k}+\alpha_k}{\sum_{v'} n_{k,v'}+\beta_{v'}} \qquad (2)$$

$$\phi_{k,v} = \frac{n_{k,v}+\beta_v}{\sum_{v'} n_{k,v'}+\beta_{v'}} \qquad (3)$$

3.2. Similarité entre items

Cette étape intègre les résultats fournis par LDA pour trouver la similarité entre items afin de prédire les préférences de l'utilisateur courant dans le filtrage collaboratif. L'estimation obtenue Θ est la distribution de thèmes latents pour chaque item, vu comme une matrice de similarité d'items par thème, et permet de calculer la similarité entre items. Chaque item possède sa propre distribution à partir

de Θ.Pour mesurer la similarité entre deux items, différentes mesures peuvent être utilisées. Nous avons choisi d'utiliser le coefficient de corrélation de Pearson car d'après [Herlocker *et al.*, 2002], en général, les résultats sont meilleurs. Chaque item est représenté comme un vecteur de thèmes et le coefficient de corrélation entre deux items i et j ayant chacune une variance (finie), noté Cor (i, j) est défini par :

$$\text{Cor (i, j)} = \frac{Cov(i,j)}{\sigma_i \sigma_j} \qquad (4)$$

Où Cov (i, j) désigne la covariance de deux items i et j, σ_i et σ_j leurs écarts types. Le coefficient de corrélation est symétrique et prend ses valeurs entre -1 et +1

3.3. Prédiction des préférences de l'utilisateur

Notre approche recommande des items pour lesquels les distributions de thèmes des titres sont similaires au profil utilisateur. L'objectif est de prédire les préférences de l'utilisateur courant aux items non consultés. Supposons que nous ayons l'historique des préférences des utilisateurs vus comme une matrice M, qui est la matrice d'évaluation employée dans le filtrage collaboratif. Nous regardons ensuite l'ensemble des items que l'utilisateur courant a déjà consulté et déterminons la similarité avec les autres items que l'utilisateur courant n'a pas encore vus en utilisant la matrice de similarité de l'étape précédente. En effectuant cela, les notes des items pour l'utilisateur courant peut être obtenue et servira à indiquer le degré de préférence de l'utilisateur courant pour les nouveaux items non consultés.

La prédiction des préférences $P_{u,\,i}$ pour un item i, pour l'utilisateur u, est basée sur la moyenne pondérée des préférences et des scores de similarité à partir de tous les autres items qui ont été notés par l'utilisateur u.

La formule est la suivante :

$$\text{Pu,i} = \sum_{j \in J} W_{u,j} * sim(i,j) \qquad (5)$$

Où J est l'ensemble des items les plus similaires à l'item i et que l'utilisateur u a noté ; $W_{u,\,j}$ est le score donné par u pour l'item $j \in J$; sim (i, j) la similarité entre les items i et j. La somme est calculée à partir de tous les items $j \in J$ notés par l'utilisateur u.

Les préférences calculées précédemment sont ordonnées par prédiction de pertinence décroissante et les N premières recommandations non notées par l'utilisateur courant sont recommandées. Notre approche a l'avantage de pouvoir prendre en compte l'oubli en considérant J non pas comme l'ensemble de tous les items déjà consultés mais les k derniers ou l'ensemble des items de la session courante, ou l'ensemble des items consultés dans la semaine courante, l'ensemble des items dans une catégorie donnée, etc. Nous laissons toutefois cette adaptation

pour des travaux futurs. Dans la section 4, l'évaluation considère J comme l'ensemble des items déjà consultés par l'utilisateur.

4. Evaluation et résultats

4.1 Cadre d'évaluation : collection, mesures et références

Pour évaluer notre méthode, nous avons utilisé la méthode de validation utilisant 90% des données pour l'entrainement du modèle et 10% de données pour le test. Pour un utilisateur dans les données tests, nous tirons aléatoirement un item supposé être en cours de consultation. A partir de cet item, le système entrainé propose les items recommandés. Si un item recommandé est effectivement noté positivement (la note est supérieure ou égale à 5 dans l'intervalle de 1 à 10) par l'utilisateur dans la collection, la recommandation est considérée comme pertinente. Nous nous appuyons pour l'évaluation sur la collection Book-Crossing et des mesures d'évaluation présentées ci-dessous.

4.1.1 Collection

Source :http://www2.informatik.uni-freiburg.de/~cziegler/BX/

Cette collection a été collectée par Cai-Nicolas Ziegler (via une exploration automatique du web) pendant quatre semaines (en 2004) à partir de la communauté Book-Crossing avec l'autorisation de Ron Hornbaker (Humankind Systems). Elle contient 278 858 utilisateurs (rendus anonymes mais avec des informations démographiques) fournissant 1 149 780 évaluations (explicites / implicites) sur environ 271 379 livres.

La collection **Book-Crossing** est constituée de trois parties :

- **BX-Users**
  contient des informations sur les utilisateurs. Les identifiants des utilisateurs ('ID-Utilisateur') ont été rendus anonymes. Des données démographiques comme ('Localisation','Age') sont parfois fournies.

- **BX-Books**
  Les livres sont identifiés par leur ISBN (International Standard Book Number) ou numéro international standard des livres. Certaines méta-données comme ('Titre du livre', 'Auteur du livre', 'Années de Publication', 'éditeur') sont fournies.

- **BX-Book-Ratings**
  contient les informations de notations du livre. Les notes sont soit explicites, exprimées sur une échelle de 1 à 10 (valeurs plus élevées indiquant une appréciation plus élevée), soit implicites, exprimées par 0.

Dans nos expérimentations, nous nous sommes focalisés sur les données pour lesquelles une notation explicite était présente ; 397 247 notations respectent cette contrainte.

4.1.2 Mesures

En recherche d'information, la précision mesure la proportion de documents pertinents dans l'ensemble de documents restitués. Dans le cas d'un SR, cette mesure peut être adaptée en la proportion d'items pertinents recommandés dans l'ensemble des items recommandés.

$$\textbf{Précision} = \frac{\boldsymbol{RP(i)}}{\boldsymbol{R(i)}} \qquad (6)$$

Où

- RP(i) est le nombre d'items recommandés et pertinents pour l'item i et
- R(i) est le nombre de documents recommandés pour l'item i.

De la même façon, nous pouvons adapter la mesure de précision moyenne pour une requête définie en recherche d'information en la précision moyenne pour un item donné :

$$\text{AP(i)} = \frac{\sum_{r=1}^{R(i)}[P@r(i).rel(r)]}{P(i)} \quad (7)$$

Où :

- P(i) est le nombre d'items recommandés et pertinents pour l'item i ;
- R(i) est le nombre d'items recommandés pour l'item i ;
- r est le rang ;
- P@r(i) est la précision lorsque les r premiers items sont recommandés pour l'item i ;
- rel(r) vaut 1 si la recommandation au rang r est pertinent et 0 sinon.

La moyenne des précisions moyennes (Mean Average Precision ou MAP) est alors la moyenne arithmétique des précisions moyennes sur l'ensemble des items considérées.

$$\text{MAP} = \frac{\sum_{i=1}^{I} AP(I)}{I} \quad (8)$$

Avec I le nombre d'items à partir desquels on recherche les items à recommander.

Ces mesures considèrent deux niveaux de pertinence : un item est soit pertinent, soit non pertinent pour un item de départ donné. Par ailleurs, ces mesures sont orientées vers l'utilisateur qui souhaite d'abord des items pertinents ; le rappel est donc moins important dans les SR.

4.1.3 Méthodes de référence

Nous avons comparé les résultats de notre méthode à 3 modèles de référence.

TFIDF (Term Frequency-Inverse Document Frequency) [Salton, 1989] est une méthode de pondération de termes qui peut être incorporée dans un système de filtrage collaboratif pour trouver la similarité entre items. Il s'agit donc d'une approche hybride.

UBCF (User Based Collaborative Filtering) ou Filtrage Collaborative Basé Utilisateur est une approche qui, à partir d'un utilisateur courant u, recherche les utilisateurs qui sont similaires à cet utilisateur en fonction de la similarité des notes qu'ils ont fournies sur les items. UBCF recommande les items que ces utilisateurs similaires ont aimés [Ekstrand *et al.*, 2011].

IBCF (Item Based Collaborative Filtering) ou Filtrage Collaborative Basé Item est une approche obtenue par la transposition de la matrice de similarité de la méthode UBCF. Alors que UBCF génère des prédictions basées sur les similaritésentre les utilisateurs, IBCF génère des prédictions basées sur les similarités entre les items [Sarwar *et al.*, 2001].

4.2 Résultats et discussions

Le but de l'expérimentation est de comparer notre méthode hybride CBCF avec une autre méthode hybride TFIDF et avec deux autres non hybrides IBCF et UBCF. Nous utilisons la collection Book-Crossing présentée plus haut.

Les figures 1 et 2 montrent une comparaison des performances des différentes méthodes. Nous rapportons la précision et la MAP en fonction du nombre d'items recommandés. Ces résultats correspondent à une moyenne de précision obtenue en prenant 50 items initiaux de différents utilisateurs choisis aléatoirement à partir desquels le système propose des recommandations.

Nous obtenons des résultats MAP ~ 0.5 et précision ~0.6 ce qui correspond à de bons résultats par rapport aux autres méthodes de la littérature.

Dans figure 1, nous obtenons la précision en fonction du nombre d'items recommandés et que nous faisons varier de 5 à 20 items.Pour toutes les méthodes nous obtenons la meilleure performance de précision en utilisant 5 items recommandés et inversement pour 20 items recommandés. Pour 5 items recommandés, notre méthode CBCF enregistre un taux de performance de 58% qui est supérieur à celui de TFIDF de 39%. IBCF et UBCF enregistrent des précisions inférieures de 9% et 6% respectivement.

Dans figure 2, nous comparons la performance de la MAP des différentes approches. Nous observons le même type de résultats que dans la figure 1. CBCF a le meilleur taux avec 47% suivi de TFIDF de 26%. IBCF et UBCF ont des taux de 7% et 5% respectivement.

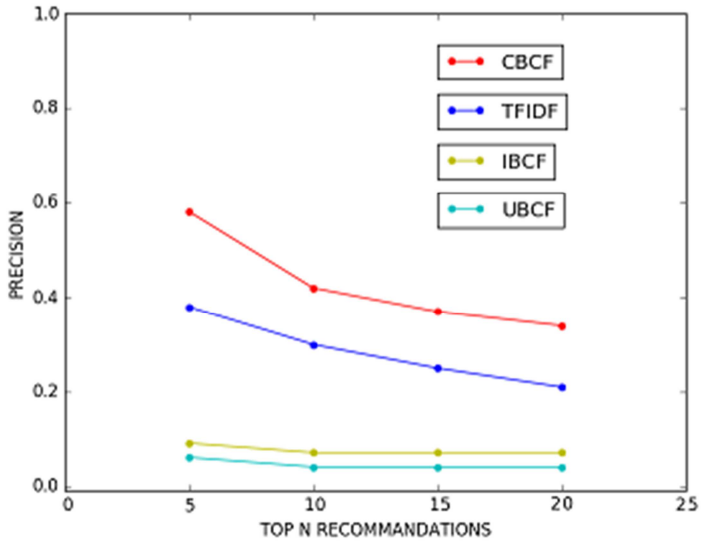


Figure 1. Précision en fonction du nombre d'items recommandés-Moyenne sur 50 items de départ utilisant différentes approches. CBCF correspond à la méthode que nous présentons.

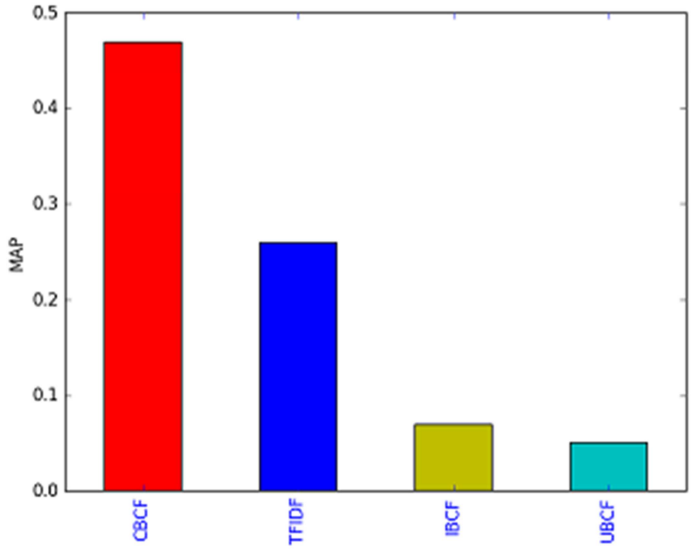


Figure 2. Comparaison de performance de la MAP-Moyennes sur 50 items (Axe des Y) de départ en utilisant les différentes approches, CBCF, TFIDF, IBCF et UBCF (Axe des X).

Dans figures 1 et 2, nous observons que les deux méthodes IBCF et UBCF utilisant des données explicites (notes) sont moins performantes. Ceci est certainement dû au fait que le filtrage collaboratif pur n'arrive pas à gérer le problème de démarrage à froid dans les deux méthodes.

CBCF et TFIDF hybride montrent de meilleure performance comparée à l'approche filtrage collaboratif grâce à leur propriété hybride et l'utilisation des données implicites (titres des livres) pour trouver la similarité entre items. Néanmoins, la méthode que nous proposons a de meilleure performance que TFIDF.

5. Conclusion

Dans cet article, nous avons proposé un SR hybride combinant LDA et la méthode de filtrage collaboratif pour des données implicites. LDA permet de trouver la structure sémantique latente dans les titres des items (ici des livres) consultés, la distribution des mots sur les thèmes latents et le mélange des distributions des thèmes latents. Le résultat provenant de LDA est ensuite intégré dans un système de filtrage collaboratif basé sur la similarité des utilisateurs.

Basée sur le thème de l'intention de l'utilisateur défini comme le profil utilisateur, notre approche recommande des items pour lesquels les distributions de thèmes des titres de l'item courant est similaire au profil utilisateur courant.

Nous avons montré que notre approche donne de meilleurs résultats par rapport au modèle hybride TFIDF. Nous devons maintenant confronter notre modèle à d'autres modèles de la littérature.

Par ailleurs, dans notre implémentation de SR sensible aux contextes, nous avons induit le contexte (intention de l'utilisateur) à partir des titres des livres. Cependant d'autres caractéristiques (auteurs, éditeurs, notes, …) des livres peuvent être utilisées pour induire d'autres contextes non observables et ainsi avoir de meilleure performance. De plus, des travaux additionnels peuvent être effectués en ajoutant des contextes observables comme la localisation, la compagnie, la période, etc.

Bibliographie

Adomavicius, G., Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering, 17*(6), 734-749.

Adomavicius, G., Tuzhilin, A. (2011). Contextaware recommender systems. In *Recommender systemshandbook*, pages 217-253. Springer.

Blei, D. M., Ng, A. Y., Jordan, M. I. (2003). *Latent dirichlet allocation. Journal of machine Learning research*, 3(Jan) :993-1022.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55(4) :77-84.*

Borras, J., Moreno, A., Valls, A. (2014). Intelligent tourism recommender systems : A survey. Expert Systems with Applications, 41(16) :7370-7389.

Burke, R. (2007). Hybrid web recommender systems. *In The adaptive web, pages 377-408. Springer.*

Candillier L., Chevalier M., Dudognon D., Mothe J. (2012). *Multiple Similarities for Diversity in Recommender Systems.International Journal On Advances in Intelligent Systems*, International Academy, Research and Industry Association, 5(3&4) :234-246.

Candillier L., Chevalier M., Dudognon D., Mothe J. (2011). *Diversity in Recommender Systems: Bridging the gap between users and systems (regular paper).International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services (CENTRIC 2011)*, p. 48-58.

Canut M.-F., On-At S., Péninou A., Sèdes F. (2015). *Enrichissement du profil utilisateur à partir de son réseau social dans un contexte dynamique : application d'une méthode de pondération temporelle. INFormatique des Organisations et Systemes d'Information et de Decision (INFORSID 2015)*, p. 15-30.

Chaker H., Chevalier M., Tricot A. (2013). *Une approche de gestion de contextes métiers pour l'accès à l'information (regular paper). INFormatique des Organisations et Systemes d'Information et de Decision (INFORSID 2013)*, p. 115-130.

Chevalier, M., Dudognon, D., Mothe, J. (2016). ADORES : a diversity-oriented online recommender system. *In Proceedings of the 31st Annual ACM Symposium on Applied Computing, pages 1075-1076. ACM.*

Dey, A. K. (2001). Understanding and using context. Personal and ubiquitous computing, 5(1) :4-7.

Ekstrand, M. D., Riedl, J. T., Konstan, J. A., et al. (2011). Collaborative filtering recommender systems. *Foundations and Trends R in Human– Computer Interaction*, 4(2) :81–173.

Fei-Fei, L. & Perona, P. (2005). A bayesian hierarchical model for learning natural scene categories. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), volume 2, pages 524-531. IEEE.

Griffths, T. L. & Steyvers, M. (2004). Finding scientific topics. Proceedings of the National academy of Sciences, 101(suppl 1) :5228-5235.

Herlocker J., Konstan J. A., Riedl J., "An empirical analysis of design choices in neighborhood-based collaborative filtering algorithms," *Information Retrieval*, vol. 5, no. 4, pp. 287–310, 2002.

Ionescu, R. T., Chifu, A. G., Mothe, J. (2015, September). DeShaTo: describing the shape of cumulative topic distributions to rank retrieval systems without relevance judgments. In *International Symposium on String Processing and Information Retrieval* (pp. 75-82). Springer International Publishing.

Lamsfus, C., Alzua-Sorzabal, A., Martin, D., Salvador, Z., and Usandizaga, A. (2009). Human-centric ontology-based context modelling in tourism. *In KEOD, pages 424-434.*

Louëdec, J., Chevalier, M., Mothe, J., Garivier, A., and Gerchinovitz, S. (2015). A multiple-play bandit algorithm applied to recommender systems. *In FLAIRS Conference, pages 67-72.*

Louëdec, J., Chevalier, M., Garivier, A.,Mothe, J. (2015), *Algorithmes de bandits pour la recommandation à tirages multiples.Document numérique*, Hermès, 18(2&3) :59-79.

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, 20(11) :1535-1549.*

Pazzani, M. J., Billsus, D. (2007). Content-based recommendation systems. *In The adaptive web, pages 325-341. Springer.*

Ramage, D., Hall, D., Nallapati, R., Manning, C. D. (2009). Labeled lda : A supervised topic model for credit attribution in multi-labeled corpora. *In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing : Volume 1-Volume 1, pages 248-256. Association for Computational Linguistics.*

Rakotonirina A. J. (2017). Filtrage Collaboratif Sensible au Contexte : une approche basée sur LDA, thèse de Master..

Ryan, N., Pascoe, J., Morse, D. (1999). Enhanced reality _eldwork: the context aware archaeological assistant. *Bar International Series, 750 :269-274.*

Salton, G. (1989). Automatic text processing : The transformation, analysis, and retrieval of. Reading : Addison-Wesley.

Sarwar, B., Karypis, G., Konstan, J., Riedl, J. (2001). Itembased collaborative filtering recommendation algorithms. *In Proceedings of the 10th international conference on World Wide Web, pages 285–295. ACM.*

Schein, A. I., Popescul, A., Ungar, L. H., Pennock, D. M. (2002). Methods and metrics for cold-start recommendations. *In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 253-260). ACM.*

Steyvers, M., Smyth, P., Rosen-Zvi, M., Gri_ths, T. (2004). Probabilistic author-topic models for information discovery. *In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 306-315. ACM.*

Wei, X., Croft, W. B. (2006). Lda-based document models for ad-hoc retrieval. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pages 178-185. ACM..

Yu, K., Zhang, B., Zhu, H., Cao, H., Tian, J. (2012). Towards personalizedcontext-aware recommendation by mining context logs through topic models. *In Pacific-Asia Conference on Knowledge Discovery and Data Mining, pages 431-443. Springer.*

Large scale reverse image search

A method comparison for almost identical image retrieval

Mathieu Gaillard[1], Előd Egyed-Zsigmond[1,2]

1 Université de Lyon
INSA Lyon,
mathieu.gaillard@insa-lyon.fr
2.Université de Lyon
LIRIS, INSA Lyon
elod.egyed-zsigmond@insa-lyon.fr

RESUME. Dans ce papier nous présentons une étude comparative de méthodes pour un système de recherche d'images inversée. Nous nous concentrons plus spécifiquement sur le cas de la recherche d'images quasi identiques dans de très grands ensembles d'images. Après une étude de l'état de l'art, nous avons implanté notre propre système de recherche d'images inversée en utilisant trois descripteurs basés sur des techniques de hachage perceptuel choisies pour leur extensibilité. Nous avons comparé la vitesse et la précision/rappel de ces méthodes contre plusieurs modifications (flou gaussien, redimensionnement, compression, rotation, recadrage). Nous proposons également un système à deux couches combinant : une première étape très rapide mais moyennement précise ; avec une étape, certes, plus lente mais beaucoup plus précise. Nous améliorons ainsi la précision globale de notre système tout en conservant sa rapidité de réponse.

ABSTRACT. In this paper, we presented our study and benchmark on Reverse Image Search (RIS) methods, with a special focus on finding almost similar images in a very large image collection. In our framework we concentrate our study on radius (threshold) based image search methods. We focused our study on perceptual hash based solutions for their scalability, but other solutions seem to give also good results.

We studied the speed and the accuracy (precision/recall) of several existing image features. We also proposed a two-layer method that combines a fast but not very precise method with a slower but more accurate method to provide a scalable and precise RIS system.

MOTS-CLES : recherche d'images inversé, pHash, optimisation, SI images

KEYWORDS: reverse image search, pHash, optimization, image information systems

1. Introduction

In this paper we deal with the reverse image retrieval problem in image centered information systems. The purpose of the reverse image retrieval is to retrieve images by similarity based on a query image. This study is mainly motivated by the need to find the original of an image in a large-scale image database given a slightly modified version of it. This kind of systems are extensively used in the context of Intellectual property and crime prevention. Many implementations already exist, for example Google Images (Google, 2017), TinEye (TinEye, 2017) and Microsoft PhotoDNA(Microsoft, 2017). At last we also expose a benchmark that we designed in order to evaluate and compare these systems.

In the introduction we present and define the problem. In the second section we will explain how reverse image search engines generally work. In the third section we will specially present the perceptual hashes as a technical solution to address our problem. We will also introduce different existing implementations. Finally, in the fourth section, we will expose our benchmark for these systems and the result of our experimentations.

In this document we study a reverse image search engine that is capable of indexing a huge amount of images and then allows the user to search for the original version of an image, even if this one is slightly modified. The number of indexed images can be more than a million. The time to index does not really matter as long as it is done within a reasonable interval. On the contrary the time to search should be as short as possible in such a way that it is possible to search 10 000 images in less than an hour. If several technical solutions are able to address this problem, we want to be able to compare them in order to select the one that best fits our needs.

We define a modification as an operation that does not alter the essential content of an image (Zauner, 2010) A non-exhaustive list of modifications is given later in this section.

The definition of a similar image varies depending on what photometric and geometric variations are deemed acceptable. This depends on the application (Philbin, Isard, & Zisserman, 2007). For our application, two images are considered perceptually similar if a human interprets and understands them as the same image. For example, an image and a slightly modified version of it are perceptually similar whereas two images with similar colors and shapes can be conceptually different therefore perceptually different.

In this section a non-exhaustive list of image modifications is exposed: Scaling: not necessarily with the same aspect ratio ; Lossy compression: JPEG ; Filter: blur, noise, artistic color filters ; Rotation ; Flipping: horizontally or vertically ; Cropping ; Shifting ; Random deletion: rectangle ; Random insertion: text, watermark, logo ;

After a modification by one of those listed above, an image can be considered as similar to the original one. These modifications do not alter the essential content of images if they are done carefully. Obviously, it is no more the case if the parameters take extreme values. For example, if we crop the half of an image, this one becomes

perceptually different. On a computer to perform these modifications on a collection of images we can use ImageMagick. (ImageMagick, 2017)

2. State of the art

Reverse image search (RIS) is a type of content-based image retrieval (CBIR). According to (Chutel & Sakhare, 2014), It is a search engine technology that takes images as input and returns results related to the query image. The search analyses the actual content of images rather than the metadata such as keywords or descriptions associated with them. The aim of Reverse Image Search Engine is to find the similar, exact image on web based on the given query image though the search images are cropped, transformed or it may have illumination changed. This Reverse Image Search engine can be used for detecting unauthorized use of brands and copyright images. Other common usage modes are to locate the source of an image, find a higher resolution version, discover other webpages where the image appears or get some more information about the image. (TinEye, 2017)

The following is a list of terms related to the reverse image search topic: near identical image detection, image retrieval, content-based image retrieval, information retrieval.

In what follows, we present our general framework for reverse image search.

Reverse image search with large databases imposes two challenging constraints on the methods used. Firstly, for each image, only a small amount of data (a fingerprint) can be stored; secondly, queries must be very cheap to evaluate. In order to be able to deal efficiently with millions of images, while still being able to keep a sizeable portion of the data in main memory, we need to generate an extremely compressed feature vector for each image. (Philbin et al., 2007)

Most approaches to reverse image search share a similar pattern. Firstly, an image representation and a distance measure are defined, which affects both the amount of data stored per image and the time complexity of a database search. When searching the database for similar images, algorithms of different time complexity are used, the most naive approach being computing the difference to every image in the database. (Philbin et al., 2007)

Reverse image search engines usually work in two phases: indexing and searching. In the indexing phase, the database is filled with feature vectors of images that should be found later on. The images are not necessarily stored in the database; this reduces the size of the database dramatically. In the searching phase, a new image is presented to the system and a feature vector representing this one is computed. Then the feature vector is compared to those in the database using the previously defined distance measure. Here are two manners to handle the results: a radius (threshold) based and a k nearest neighbors based method. The radius based method works well for content authentication emphasizing precision while the k nearest neighbors, usually is used when the needs of recall are more important. A general workflow for the radius based method is illustrated in Figure 1. There is a match if the distance between the query

vector and a vector representing an indexed image is less than or equal to a threshold. Usually the distances are normalized between 0 and 1. If there is a match the system will return the images corresponding to the concerned feature vectors as results to the similar image search for the query image. There are a lot of techniques to define a pair of image representations and distances. Some of them are discussed in the next section. The workflow is composed of these steps:

Feature vector extraction: A feature vector is computed from the image using one of the later-discussed techniques.

Matching: A feature vector is compared to those in the database. A sequential search is the easiest way to iterate over the database and can be parallelized or even distributed among many computers. There also exist some special data structures to fasten the search when the distance verifies certain properties. Instead of considering all indexed images, only a subset of them is compared to the query feature vector. For example, in Hamming space, the Multi-Index Hashing (MIH) has sub-linear run-time behavior for uniformly distributed codes (Norouzi, Punjani, & Fleet, 2013).

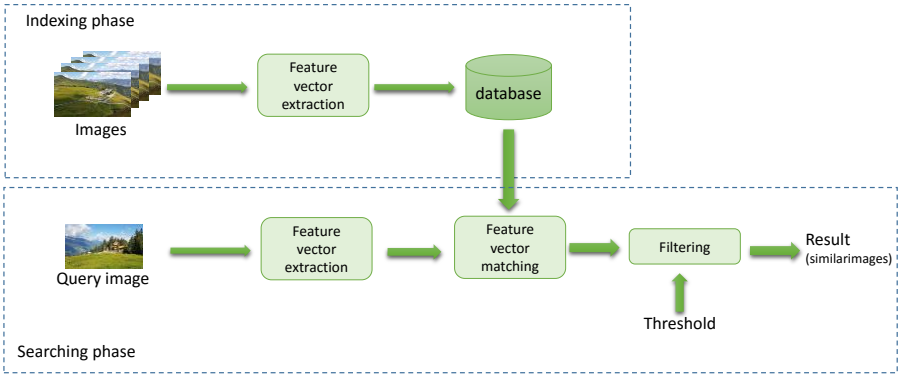


Figure 1 General workflow for reverse image search.

Therefore the determination of an adequate threshold, in accordance with the actual application scenario, is critical. (Zauner, 2010) Information retrieval research has shown that precision and recall follow an inverse relationship. (Datta, Joshi, Li, & Wang, 2008) If the threshold is too low, the precision is better at the expense of the recall because only the most relevant images are retrieved. On the contrary, if the threshold is too high, the recall is better but the precision is worse. When the application needs to authenticate an image, the precision is more important because we want to limit the number of false positives, whereas, when the application needs to identify content, the recall is more important because the user can deal with a small number of false positives. (Zauner, 2010) In any case, relevant and irrelevant images cannot be separated clearly; the boundary between these two sets is fuzzy.

3. Technical solutions

As stated in the last section, a technical solution for reverse image search consists of an image representation and a distance measure between these representations. (Philbin et al., 2007) In addition to that, the image representation should be easy to store in a database. For example, to compute a similarity distance between two images first their feature vectors are extracted and then the distance between them is measured. The more the result is near to 0, the more the images are considered as similar. In our study we have considered several image features based on colors, textures, contours, interest points, neural networks and tested libraries such as LIRE[1](Lux & Chatzichristofis, 2008), OpenCV[2] and pHash[3]. We focused our study on perceptual hashing because, at a first glance, it appeared to be significantly faster than the descriptors from the LIRE library with an almost identical accuracy, on basic modifications (scaling, compression, grayscale filter).

As our goal was to search large image collections, we focused mainly on the perceptual hash functions implemented in the library pHash because it was really faster than the others giving comparable accuracy. Moreover, by using an existing library we save implementing time in order to focus our resources on experimentations.

3.1. Perceptual hashing

3.1.1 Definition

A perceptual hash function is a type of hash function that has the property to be analogous if inputs are similar. This allows us to make meaningful comparisons between hashes in order to measure the similarity between the source data. Perceptual hash functions are an interdisciplinary field of research. Cryptography, digital watermarking and digital signal processing are part of this field of research.

The definition of a hash function according to (Menezes, Oorschot, & Vanstone, 1997) is:

> *A hash function is a computationally efficient function mapping binary strings of arbitrary length to binary strings of some fixed length, called hash-values.*

In the case of a perceptual hash, some more properties should be present according to (Zauner, 2010):

> *Let H denote a hash function which takes one media object (e.g. an image) as input and produces a binary string of length l. Let x denote a particular media object and $\hat{x}$ denote a modified version of this media object which is "perceptually similar" to x. Let y denote a media object that is "perceptually*

[1] http://www.lire-project.net
[2] http://opencv.org
[3] http://phash.org

different" from x. Let x' and y' denote hash values. $\{0/1\}^l$ *represents binary strings of length l. Then the four desirable properties of a perceptual hash are identified as follows.*

A uniform distribution of hash-values; the hash-value should be unpredictable.

$$P(H(x) = x') \approx \frac{1}{2^l}, \forall x' \in \{0/1\}^l$$

Pairwise independence for perceptually different media objects.

$$P(H(x) = x' | H(Y) = y') \approx P(H(x) = x'), \forall x', y' \in \{0/1\}^l$$

Invariance for perceptually similar media objects.

$$P(H(x) = H(\hat{x})) \approx 1$$

Distinction of perceptually different media objects. It should be impossible to construct a perceptually different media object that has the same hash-value as another media object.

$$P(H(x) = H(y)) \approx 0$$

Most of the time to achieve these properties the perceptual hash function extract some features of media objects that are invariant under slight modifications to construct a perceptual hash. For example, knowing how a compression algorithm works, it is possible to find some invariant features and then design a perceptual hash based on them. Some examples of perceptual hash functions for images are detailed later in this section.

3.2. Implementations

In this section, we present the three implementations of perceptual hash functions that we used in our benchmark: the DCT based , the Marr-Hildert Operator based and the Radial Variance based perceptual hash functions from (Zauner, 2010).

The Discrete Cosine Transformation (DCT) based perceptual hash from (Zauner, 2010) takes advantage of the property that low-frequency DCT coefficients are mostly stable under image modifications to construct a 64 bits image hash. The Hamming distance is used to compare them. The fact that the hashes are encoded on 64 bits and the use of the Hamming distance is a wise choice because a hash can fit in a processor register. Moreover, to fasten the calculation of the Hamming distance, it is possible to use the special popcount (Sun & Mundo, 2016) instruction from the x86 processor family.

The Marr-Hildreth (MH) operator, also denoted as the Laplacian of Gaussian (LoG), is a special case of a discrete Laplace filter. It is an edge and contour detection based image feature extractor. The MH operator generates vectors encoded on 576 bits.

The Radial Variance hash (Standaert et al., 2005) is based on the Radon transform that is the integral transform which consists of the integral of a function over a straight line. It is robust against various image processing steps (e.g. compression) and more robust than the DCT and MH based perceptual hash functions against geometrical transformations (e.g. rotation up to 2°).

4. Benchmarking

As stated in (Zauner, 2010) not much research has been published dealing with the benchmarking of perceptual hash functions. Therefore, they propose their own benchmark for perceptual hash function: Rihamark. This one allows the user to compare several perceptual hash functions against several attacks and to analyze the results with graphics. The benchmark is modular so that it is possible to add new perceptual hash functions, attacks functions or analyzer functions. This benchmark is for example very useful to choose which perceptual hash function is best suited for a specific usage.

Currently we didn't find an already implemented benchmark for a reverse image search engine. It is important to have a benchmark to test all the technical solutions previously detailed. As previously said there exists a benchmark but only for perceptual hash functions. However, it could be adapted to reverse image search. In fact, (Zauner, 2010) proposes some metrics to evaluate content identification systems but no implementation is provided along with it. Their approach comes from a biometrics background because they model the search as m authentication tests. Basically they calculate the False Accept and the False Reject Rate (FAR/FRR) and then plot the Receiver Operating Characteristic (ROC) curve. They also explain how to compare several perceptual hash functions based on their respective ROC curves.

We designed a new framework to benchmark the reverse image search engines. Our approach is based on the evaluation measures of information retrieval systems described in (Manning, Raghavan, & Schütze, 2009). We model the reverse image search engine as an information retrieval system that returns an unranked set of documents for a query. If many documents are retrieved, the user is in charge of choosing the best suited image.

4.1. Metrics

4.1.1 Effectiveness

To process a query, the reverse image search engine classifies the indexed images by relevance. In addition, we introduce a threshold for each similarity measure to be able to precisely select the images that are to be considered as similar to the query image. Thus, each image, whether relevant for the query or not, can be retrieved or not. This notion can be made clear by examining the following contingency table from (Manning et al., 2009).

	Relevant	Nonrelevant
Retrieved	True positive	False positive
Not retrieved	False negative	True positive

The effectiveness of the system is measured with the following metrics from (Manning et al., 2009)

Precision (P) is the fraction of retrieved documents that are relevant.

$$Precision = \frac{\#(relevant\ items\ retrieved)}{\#(retrieved\ items)} = P(relevant|retrieved)$$

Recall (R) is the fraction of relevant documents that are retrieved.

$$Recall = \frac{\#(relevant\ items\ retrieved)}{\#(relevant\ items)} = P(retrieved|relevant)$$

A single measure that trades off precision versus recall is the F measure, which is the weighted harmonic mean of precision and recall:

$$F = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \ ; \ F_{\beta=1} = \frac{2PR}{P + R}$$

It is possible to change the weights in the harmonic mean of the F measure in order to tune it. This is done by changing the β parameter. Values of $\beta < 1$ emphasize precision, while values of $\beta > 1$ emphasize recall. This is important in order to benchmark the system in accordance with its application.

4.1.2. Performance

It is important to measure the time taken by the system for indexing and searching. Actually, the system should be able to index millions of images and search across them as fast as possible. The complexity of both the indexing and searching phases depends on the number of images and is not necessarily linear. In fact, the data structure used to store the feature vectors can have a nonlinear complexity. Therefore, we propose to measure the indexing and searching time for a certain number of images.

4.2. Protocol

In order to compare the effectiveness and the speed of different image search methods, we created a comparison protocol. We chose 25 000 images from the mirflickr dataset[4].

In order to be able to calculate automatically the precision and recall of the results, we applied 6 small modifications on each image, that gave us a dataset with 175 000

[4] Http://press.liacs.nl/mirflickr/ (consulted in 2017)

images. We measured the index and search speed as well as the results precision and recall.

The modifications (illustrated on Figure 2) applied on the images are:

1. Gaussian blur (r=4, Σ=2)
2. Black and white transformation
3. Resize to half height and width
4. Compression into jpeg with a quality=10
5. Clockwise rotation by 5°
6. Crop by 10% at the right side of the image.

Figure 2 Illustrations of the image modifications

These modifications represent the basic cases of small changes images usually undergo over the web.

The benchmark was centered on high speed image search methods. We used only one perceptual hash function to retrieve the results. Our first benchmark protocol for a generic reverse image search engine is detailed in this section.

1. Select N+M images that are representative to an application with no duplicated images. In our case N=24 000 and M=1000 (the first 1000 images from the dataset in alphabetical order)
2. Split them into 2 sets of N base images and M non-indexed images.
3. Select K transformations and from the N base images, generate K new image sets containing K*N transformed images. In our case K=6, and the transformations are those enumerated above.

4. Index the N base images and the N*K transformed images according to the different image descriptor extraction methods. For us: Discrete Cosine Transform (DCT), Marr-Hildert Operator (MH) and Radial Variance (RV) based perceptual hashes (Zauner, 2010).
5. Make search queries with:
    a. The M images from the non-indexed image set.
    b. The N images from the base image set.
    c. The K*N images from the transformed sets.
6. Analyze the search results and compute the mean precision, the mean recall and the mean F measure of all queries.
    a. For the M images of the non-indexed set, there should be no relevant result image. Thus the result should be empty.
    b. When querying with one of the other (K+1)*N already indexed images, the relevant results are the K+1 images that are the transformations of the query image. Thus the result of each query should contain K+1 images that come from the same base image as the query image.

It is possible to repeat this protocol for several different thresholds in order to choose the best one suited for an application. In order to be able to measure the precision and the recall of our image retrieval information system, we decided to apply thresholds to the similarity scores between the query and the result images. We obtained this way a precise result set with a given item count. Having a threshold to distinguish the similar and different images enables our method to be considered also as an information retrieval system that returns an unranked set of images for a query.

4.3. Results

4.3.1. Improving searching time in Hamming space

In the case of our study we implemented and benchmarked 3 solutions to search for 64 bit hashes in Hamming space: a CPU based, a GPU based and a MIH (Norouzi, Punjani, & Fleet, 2013) based solution. For a large number of hashes (at least 50M) the MIH solution is the most efficient, followed by the GPU and finally the CPU. The two latter methods are memory bound thus the memory bandwidth and cache are both performance factors.

4.3.2. Effectiveness against modifications

In order to evaluate the effectiveness of the DCT, MH and RV based perceptual hash from pHash (Zauner, 2010) against modifications, we indexed the N base images and then made K search queries each with all N modified images.

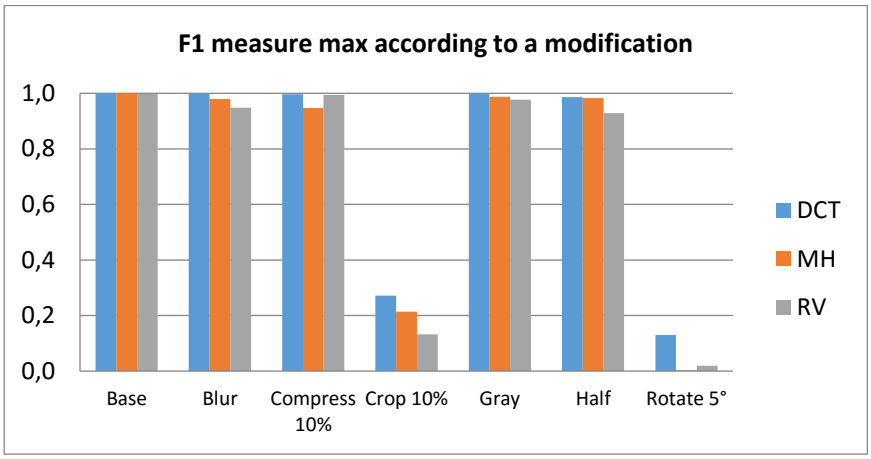


Figure 3 Maximum F1-measure of DCT, MH and RV perceptual hash function against modifications

We computed the F1-measure for various thresholds and took the maximum. The functions are robust against Gaussian blur (r=4, Σ=2), JPEG compression (quality 10%), grayscale filter, and scale to half the size. However the functions are not robust against crop (10% on the right) and rotate (5° clockwise) modifications as illustrated on Figure 3.

4.3.3. One-layer system

In a first implementation, we tested the speed and accuracy of a Reverse Image Search system based on one perceptual hash function. The experiment was carried out in order to get the best threshold values for the different methods (DCT (Figure 4), MH (Figure 5) and RV (Figure 6)).

We implemented the protocol as Linux shell commands and C++, using processor based and GPU based optimizations based on (Sun & Mundo, 2016) and other online available libraries[5].

The first results (see Table 1) showed that the DCT based hash was clearly faster than the Marr-Hildert Operator and Radial Variance based hash. It was also more accurate against the 6 chosen modifications. We tested the descriptor accuracy, calculating the mean precision, recall and F-measure of the returned results. We carried out the calculations for different threshold values. The threshold here is a normalized descriptor distance value, below which two images are considered as similar.

[5] See our implementation on : https://github.com/mgaillard/pHashRis

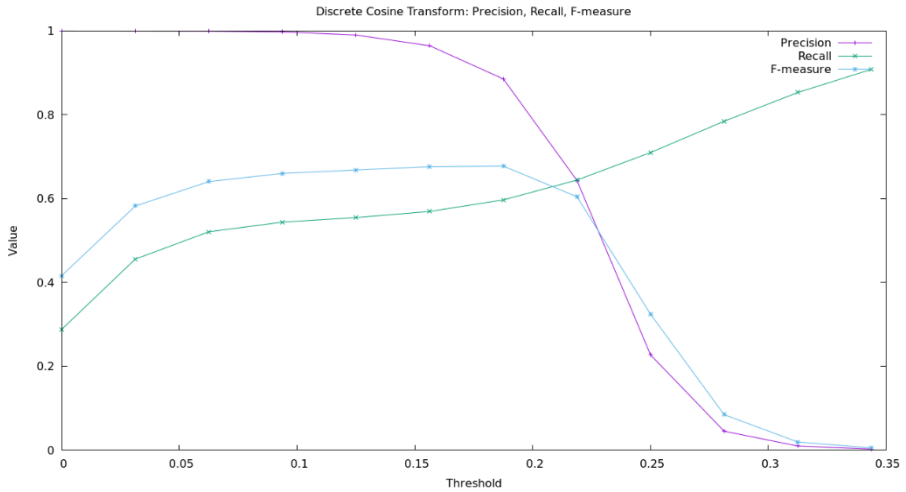


Figure 4 Precision/Recall/F-measure curves of the DCT based perceptual hash function search results according to different threshold values

The performance is evaluated through the index and search speed. Table 1 presents the measurements done for 125 000 images on an OVH dedicated virtual machine equipped with an Intel Xeon Haswell 8 cores at 3.1 GHz and 30 GB of RAM.

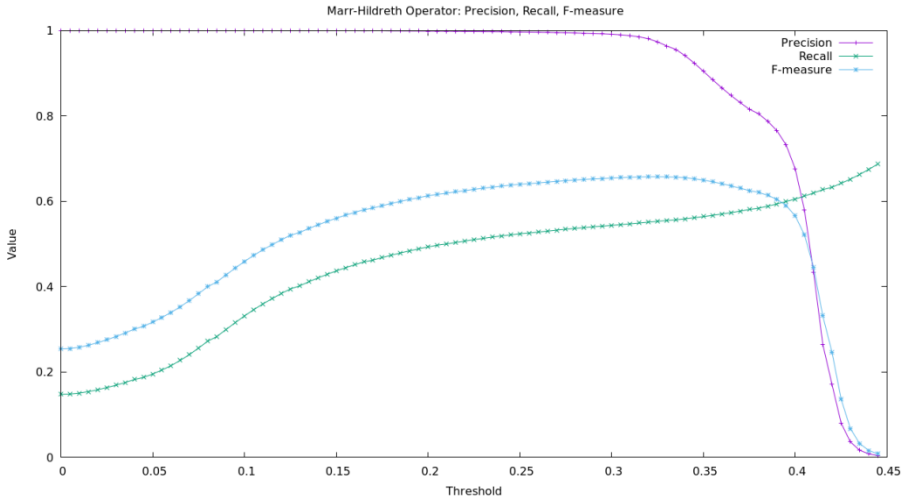


Figure 5 Precision/Recall/F-measure curves of the Marr-Hilderth based perceptual hash function search results according to different threshold values.

We also tested, the evolution of our accuracy measures with different degrees of the modifications. We noticed that the different hash methods were quite sensible to rotations (above 2°) or to cropping (above 5%), but resisted very well to compression, blur or resize.

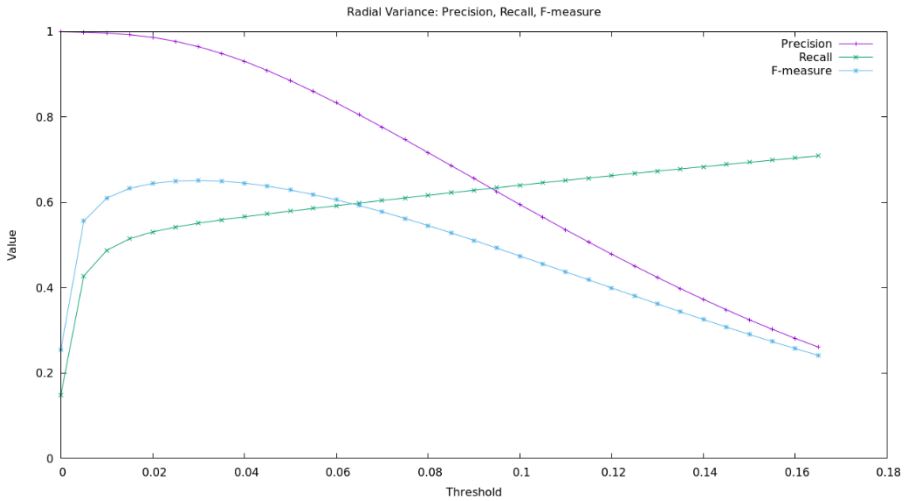


Figure 6 Precision/Recall/F-measure curves of the Radial Variance based perceptual hash function search results according to different threshold values.

	Index 125k images	125k queries on 125k images
DCT	3 min 55 sec	3 min 59 sec
MH	31 min 13 sec	32 min 20 sec
RV	1 min 48 sec	3h 47 min 31 sec

Table 1, image indexing and search time measures for the DCT, MH and RV based perceptual hash methods.

4.3.4. Two layered system

A second experiment was carried out in order to enhance the precision of our reverse image search engine. We used two successive layers of reverse image search. First, a very fast layer whose recall is high and precision is not perfect based on one perceptual hash function. Second a more accurate layer whose precision and recall are both near to 1. The aim of the first layer is to drastically reduce the number of images to be processed by the second layer which is more accurate but more expensive in searching time. The second layer is based on SIFT descriptors and on a distance between them. The distance is defined as follows:

When comparing two images A and B, let I be the number of interest points in image who has less interest points and J be the number of interest points in other image. We compute a matching that gives us for each interest point of image A the 2 nearest neighbors in the image B. Among these matches, we select only the G good matches that pass the ratio test proposed by (Lowe, 2004) in section 7.1 (In our

implementation 0,8). Finally, the distance is D = 1 – G/I. The two images are considered similar if the distance is less than or equal to 0,9.

In order to compare the results with or without the second, SIFT based layer, we run the protocol first without it in order to estimate a reasonable threshold for the perceptual hash layer and then with the SIFT layer in order to be able to compare the results. We choose the threshold for the perceptual hash layer so that the average precision of the first layer is around 0.4 thus the SIFT layer has to deal with 20 images on average which is reasonable. Because the search time is higher, we run it with these parameters: N=2000, M=200.

The results of the comparisons between the precision and recall evolutions in a one layer (DCT based perceptual hash) and a two layers (DCT based perceptual hash and SIFT based descriptor) information retrieval system are presented in Figure 7.

The precision is enhanced without affecting too much the recall at the expense of searching time and index size. We can see that for a threshold of around 0.25, the precision of the first step DCT perceptual hash based method is around 0.7. That means that in our case, when 7 images are to be returned, around 15 images are retrieved. The SIFT based comparison, even in its brute force implementation runs very fast on such a small number of images and enables to improve the global precision considerably (to more than 0,95). The implementation of our benchmarks can be accessed on GitHub[6].

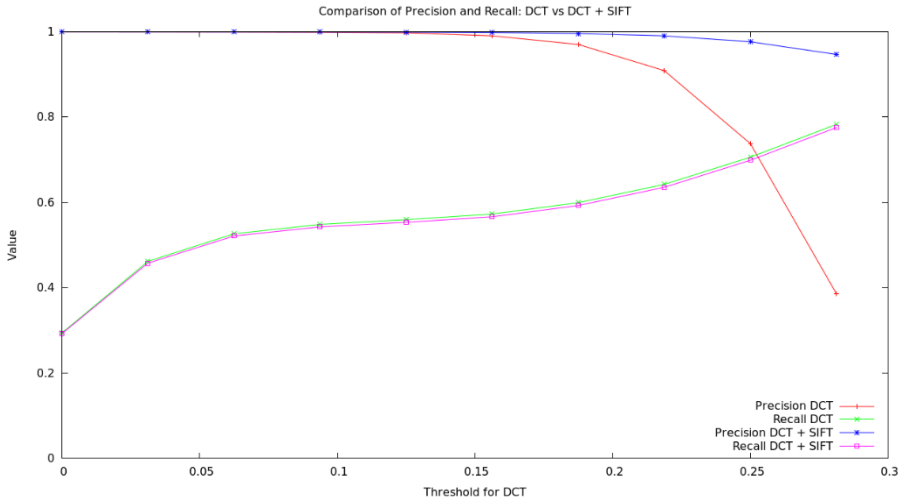


Figure 7 Precision and recall evaluation comparison between a one layer (DCT) RIS system and a two layers (DCT + SIFT) RIS system

[6] https://github.com/mgaillard/pHashRis

5. Conclusion

In this paper, we presented our study and benchmark on Reverse Image Search (RIS) methods, with a special focus on finding almost similar images in a very large image collection.

In our framework for reverse image search we state that there are mainly two algorithms for the search phase. One searches for the k nearest neighbors and the other searches for all images in a radius (within a distance threshold). We studied more the latter so the metrics of the benchmark should be adapted in the case of a k nearest neighbor method because the results can no longer be modeled as an unranked set of documents.

We focused our study on perceptual hash based solutions for their scalability, but other solutions seem to give also good results.

We studied the speed and the accuracy (precision/recall) of several existing image features. We also proposed a two-layer method that combines a fast but not very precise method with a slower but more accurate method to provide a scalable and precise RIS system.

The two-layer method can be extended with multiple functions in each layer. It is especially possible to use several perceptual hash functions in the first layer each of them tailored to a special modification. For example, a DCT based perceptual hash is robust against JPEG compression, Gaussian blur, scaling but is weak against rotation. To compensate for this weakness we could use a Color Moment based perceptual hash (Tang, Dai, & Zhang, 2012) which is robust against rotations. The second, SIFT based, layer has also some weaknesses. For example, images on which colors are uniform have a small number of interest points. To address this problem, it could be relevant to use other more accurate methods.

We foresee also the implementation of the LSH method from (Philbin et al., 2007) to reduce even more the hash sizes. The test of neural network based methods, such as the DSRH method from (Yao, Long, Mei, & Rui, 2016) is also a possible improvement idea although it needs a quite heavy learning phase.

We implemented our method in a near duplicate image search application[7] that can detect near duplicate image groups very quickly. This application can integrate information systems containing many images. One of our application fields is the illegal copy detection in large image sets. In this situation, the application has two inputs: the original images and the image set to check to search their copies in. The output will provide for each original image the list of its near duplicates found in the images set to check.

[7] https://github.com/mgaillard/ImageDuplicateFinder

References

Chutel, P. M., & Sakhare, A. (2014). Evaluation of compact composite descriptor based reverse image search. In *2014 International Conference on Communication and Signal Processing* (pp. 1430–1434). IEEE. http://doi.org/10.1109/ICCSP.2014.6950085

Datta, R., Joshi, D., Li, J., & Wang, J. Z. (2008). Image retrieval. *ACM Computing Surveys*, *40*(2), 1–60. http://doi.org/10.1145/1348246.1348248

Google. (2017). Google Images. Retrieved February 18, 2017, from https://images.google.com/

ImageMagick. (2017). ImageMagick @ Convert, Edit, Or Compose Bitmap Images. Retrieved February 18, 2017, from https://www.imagemagick.org/

Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, *60*(2), 91–110. http://doi.org/10.1023/B:VISI.0000029664.99615.94

Lux, M., & Chatzichristofis, S. A. (2008). Lire: lucene image retrieval. In *Proceeding of the 16th ACM international conference on Multimedia - MM '08* (p. 1085). New York, New York, USA: ACM Press. http://doi.org/10.1145/1459359.1459577

Manning, C. D., Raghavan, P., & Schütze, H. (2009). An Introduction to Information Retrieval. *Information Retrieval*, (c). http://doi.org/10.1109/LPT.2009.2020494

Menezes, A. J., Oorschot, P. C. Van, & Vanstone, S. a. (1997). Handbook of Applied Cryptography. *Electrical Engineering*, *106*, 780. http://doi.org/10.1.1.99.2838

Microsoft. (2017). PhotoDNA Cloud Service. Retrieved February 18, 2017, from https://www.microsoft.com/en-us/photodna

Norouzi, M., Punjani, A., & Fleet, D. J. (2013). Fast Exact Search in Hamming Space with Multi-Index Hashing. Retrieved from http://arxiv.org/abs/1307.2982

Philbin, J., Isard, M., & Zisserman, A. (2007). Scalable Near Identical Image and Shot Detection. *Analysis*, 549–556. http://doi.org/10.1145/1282280.1282359

Standaert, F.-X., Lefebvre, E., Rouvroy, G., Macq, B., Quisquater, J.-J., & Legat, J.-D. (2005). Practical evaluation of a radial soft hash algorithm. In *International Conference on Information Technology: Coding and Computing (ITCC'05) - Volume II* (p. 89–94 Vol. 2). IEEE. http://doi.org/10.1109/ITCC.2005.229

Sun, C., & Mundo, C. C. del. (2016). Revisiting POPCOUNT Operations in CPUs/GPUs. In *The International Conference for High Performance Computing, Networking, Storage and Analysis* (p. 2p (poster)). Salt Lake City, Utah, USA.

Tang, Z., Dai, Y., & Zhang, X. (2012). Perceptual Hashing for Color Images Using Invariant Moments. *Appl. Math. Inf. Sci*, *6*, 643–650.

TinEye. (2017). TinEye Reverse Image Search. Retrieved February 18, 2017, from https://tineye.com/

Yao, T., Long, F., Mei, T., & Rui, Y. (2016). Deep Semantic-Preserving and Ranking-Based Hashing for Image Retrieval. *International Joint Conference on Artificial Intelligence (IJCAI)*, (c), 3931–3937.

Zauner, C. (2010). *Implementation and benchmarking of perceptual image hash functions*. *Master's thesis*. Upper Austria University of Applied Sciences, Hagenberg Campus. Retrieved from http://phash.org/docs/pubs/thesis_zauner.pdf

Processus : concepts et ingénierie

Evaluation des systèmes d'information à base de technologies émergentes

Application à la blockchain

Jacky Akoka[1], Isabelle Comyn-Wattiau[2]

1. CEDRIC-CNAM et Institut Mines Telecom (TEM)
2 Rue Conté, 75003 PARIS, France
jacky.akoka@lecnam.net

2. ESSEC Business School
1 Avenue Bernard Hirsch, 95021 CERGY, France, wattiau@essec.edu

RESUME. Les technologies émergentes représentent une innovation majeure qui offre des avancées significatives aux organisations tant privées que publiques. Un exemple de ces technologies est la « blockchain » qui combine des mécanismes de cryptographie et l'architecture pair à pair. L'importance que prennent ces technologies émergentes nécessite le recours à des méthodes d'évaluation afin d'appréhender leur apport et les risques associés. L'objectif de cet article est de proposer une méthode d'aide à l'évaluation des systèmes d'information à base de technologies émergentes et de l'illustrer par l'application à la «blockchain ». Cette démarche de guidage est fondée sur le constat que les technologies émergentes sont des systèmes complexes. Notre démarche combine trois cadres conceptuels : la théorie sous-jacente aux systèmes d'information complexes, la systémique et la norme ISO 25001 dédiée à la qualité des logiciels. Nous proposons une hiérarchie multicritères qui sert de base à l'évaluation et nous l'appliquons au cas particulier de la « blockchain ».

ABSTRACT. Emerging technologies represent a major innovation that offers significant advances to both private and public organizations. An example of these technologies is the "blockchain", which combines cryptographic mechanisms and peer-to-peer architecture. The importance of these emerging technologies requires the use of evaluation methods in order to understand their contribution and the associated risks. The objective of this article is to propose a method supporting the evaluation of emerging technologies and to apply it to blockchains. This guidance approach is based on the recognition that emerging technologies are complex systems. Our approach combines three conceptual frameworks: the underlying theory of complex information systems, systems theory, and the ISO 25001 standard devoted to software quality. We propose a multi-criteria hierarchy which serves as the basis for the evaluation. To illustrate this approach, we apply it to the particular case of "blockchains".

Mots-clés : technologies émergentes, blockchain, SI complexe, méthode, évaluation, guidage.

KEYWORDS: emerging technology, blockchain, complex IS, methodology, assessment, guidance.

1. Introduction

Selon (Day *et al.*, 2004), les technologies émergentes (TE) représentent une innovation qui a le potentiel de transformer une industrie existante et/ou d'en créer de nouvelles. Bien entendu, il existe d'autres définitions. Rotolo *et al.* (2015) les résument en cinq caractéristiques principales : (i) innovation radicale, (ii) croissance rapide, (iii) cohérence, (iv) impact significatif et (v) incertitude et ambiguïté.

La différence principale entre les TE et les nouvelles technologies traditionnelles réside dans le fait que ces dernières offrent des changements incrémentaux alors que les premières sont caractérisées par une innovation radicale. Cozzens *et al.* (2010) considèrent que les TE sont celles qui possèdent un potentiel élevé mais n'ont pas encore démontré leurs valeurs ni fait l'objet d'un consensus. Song et Yin (2007) définissent quatre étapes dans l'évolution des TE. La première étape est relative à la mutation technologique, la deuxième est l'implantation des technologies, la troisième est caractérisée par l'application de l'innovation. Enfin la dernière correspond à l'innovation par l'intégration de technologies.

Il existe de nombreux exemples de TE. Le MIT présente une liste non exhaustive de ces TE par année[1]. Mentionnons les nanotechnologies qui possèdent un potentiel de création dans de nombreuses industries, comme l'industrie pharmaceutique, l'aérospatiale, l'informatique quantique, les dispositifs médicaux, l'industrie textile, etc. (Letaba *et al.*, 2014). Il existe un ensemble de ces TE particulièrement utiles pour les systèmes d'information de lutte contre l'incendie. Citons les cartes digitales spécialisées, les drones dédiés, les robots terrestres, les systèmes d'information d'urgence et les vêtements de protection intelligents (Schlauderer *et al.*, 2016). Citons aussi les différentes technologies mises en œuvre dans le « cloud » et plus particulièrement l'apprentissage mobile par le cloud (Cloud Mobile Learning) (Al-Arabiat *et al.*, 2015). Les technologies de stockage de l'énergie en sont un autre exemple, notamment celles qui transforment l'électricité en d'autres formes (chimique, mécanique) ainsi que les « smart grids » (Wang *et al.*, 2017).

La plupart des auteurs s'accordent à considérer qu'une des caractéristiques principales de ces TE est la complexité. Ainsi Huang et Yuan (2010) affirment que les TE ont un degré d'incertitude élevé et une complexité certaine. C'est cette dernière qui différencie les TE des technologies existantes. Ces TE constituent des systèmes complexes, et plus généralement des systèmes d'information (SI) complexes.

Une question importante est relative à la capacité des organisations à évaluer l'apport de ces TE et les risques associés. L'objectif principal de cet article est précisément de proposer une démarche d'évaluation des SI à base de technologies émergentes tenant compte de la complexité qui les caractérise et d'illustrer cette démarche par l'évaluation de la «blockchain», qui constitue aujourd'hui l'innovation disruptive la plus importante. Le reste de l'article est organisé de la façon suivante. Le deuxième paragraphe est consacré à un état de l'art sur la complexité, les

[1] https://www.technologyreview.com/lists/technologies/2016/

systèmes complexes, et les méthodes d'évaluation de technologies émergentes considérées comme des systèmes complexes. Notre approche de l'évaluation des technologies émergentes est présentée dans la troisième partie. La partie suivante est dédiée à l'application de cette démarche à la technologie émergente que constitue la « blockchain». Nous concluons dans la dernière partie et présentons quelques voies de recherche futures.

2. Etat de l'art

La revue de littérature synthétisée dans cette partie porte sur la notion de complexité, caractéristique importante des SI à base de technologies émergentes. La théorie des systèmes est aussi rappelée, notamment parce qu'elle intègre la complexité. Enfin, les méthodes d'évaluation des SI complexes sont résumées.

Il existe plusieurs définitions de la complexité qui reflètent les différents systèmes concernés et leurs contextes. Lloyd (2001) présente une trentaine de définitions de la complexité ainsi que les mesures qui y sont associées.

Baccarini (1996) considère que la complexité possède deux dimensions : organisationnelle et technologique. La complexité organisationnelle est définie par l'ampleur de la différenciation existant à l'intérieur des différents éléments qui constituent l'organisation. La complexité technologique est relative à la variété et à la diversité des tâches, ainsi qu'à l'interdépendance existant entre elles (Fitsilis, 2009). Geraldi et Adlbrecht (2007) caractérisent la complexité par trois dimensions : la confiance, le fait et l'interaction.

Dans la théorie des systèmes, la complexité est assimilée à un grand nombre d'éléments qui interagissent et dont le comportement individuel ne peut être prédit. De tels systèmes sont auto organisés et possèdent donc la propriété d'émergence qui permet de générer de nouvelles propriétés (Skyttner, 2005).

La théorie des systèmes complexes considère que ces systèmes sont caractérisés par leur degré d'auto organisation, par la propriété d'émergence, leur caractère innovant, leur capacité d'apprentissage et par leur pouvoir d'adaptation (Sommerville *et al*., 2012). La recherche dans ce domaine se focalise sur des notions telles que l'émergence de propriétés collectives, le comportement chaotique, l'auto-organisation, la redondance, la récursion, etc. (Holland, 2006). Certains auteurs considèrent que l'interdépendance et la taille ont un effet important sur la complexité. D'autres mettent davantage l'accent sur l'incertitude (Perminova *et al*., 2008).

Nous considérons que les technologies émergentes sont des systèmes complexes car ils possèdent toutes les caractéristiques et les attributs décrits plus haut. Leur complexité est due notamment au nombre important de composants, aux types de relations existant entre les composants, aux types de relations entre le système et son environnement, ainsi qu'à la propriété émergente de ces systèmes. Ces derniers peuvent être vus comme composés d'un grand nombre d'agents auto-organisés, qui interagissent de façon dynamique et non linéaire (Kim et Kaplan, 2006).

Plus généralement, nous les assimilons à des SI complexes qui doivent répondre rapidement à des changements dans les dimensions sociotechniques et à des besoins non fonctionnels. Ils doivent aussi prendre en compte les changements des exigences des utilisateurs, des besoins organisationnels (processus métiers et règles de gestion), des interdépendances accrues entre les individus, les organisations et les technologies. Ils doivent aussi intégrer les changements de l'environnement de ces systèmes tels que ceux des marchés, des organismes de régulation, de la concurrence, des menaces et des opportunités. Enfin, ils doivent permettre de faire face aux changements générés par les solutions propriétaires, les logiciels libres et par l'émergence de nouvelles applications et de nouveaux protocoles. Plus généralement, ils doivent résoudre les problèmes qui découlent des évolutions rapides des technologies de l'information qui constituent une dimension importante des systèmes d'information complexes.

Les changements précédents caractérisent fortement les systèmes d'information complexes, du fait notamment de :

– leur taille : ce sont souvent des systèmes de grande taille en termes de composants et sous-composants, de sites, de volume, de nature, et de rapidité de l'information.

– leurs interconnexions : en effet, le comportement du système émerge de l'interaction entre les composants, générant un comportement difficile à caractériser.

– les demandes évolutives : l'environnement dans lequel évolue le système requiert des adaptations sur une échelle de temps plus petite, comparée à celle nécessaire au développement et au déploiement.

– les technologies évolutives : les architectures matérielles et logicielles ainsi que les protocoles évoluent rapidement.

Il existe plusieurs approches de gestion de la complexité, notamment la théorie des systèmes complexes adaptatifs (Holland, 2006), la théorie réductionniste (Emmeche *et al*., 1997) et la théorie des systèmes (Skyttner, 2005). Tout comme nous avons considéré que les technologies émergentes sont des systèmes d'information complexes, nous considérons aussi que la théorie des systèmes est la plus adaptée pour faciliter la gestion de la complexité. En effet, la complexité du système est liée à sa structure, à son comportement et à sa relation avec l'environnement. Ces trois éléments sont précisément les caractéristiques principales de la théorie des systèmes (Sommerville *et al*., 2012).

(Mala et Cil, 2011) décrivent de nombreuses mesures de la complexité ainsi que leurs limitations. Les métriques généralement retenues sont fondées sur la taille du système considéré, son entropie, l'information, la hiérarchie des coûts et l'organisation. D'autres métriques sont proposées, notamment celles fondées sur les contributions de Shannon. Des exemples d'évaluation de systèmes complexes sont présentés par (Owen, 2007). Il existe de nombreuses méthodes d'évaluation des technologies. (Tran et Daim, 2008) proposent une taxonomie de ces méthodes en différenciant celles applicables au domaine public de celles qui sont utilisées dans le domaine privé. En ce qui concerne le domaine public, des méthodes fondées sur la théorie des systèmes telles que la modélisation structurelle ou la dynamique des

systèmes ont donné lieu à des techniques de type Electre, Spin ou Qsim. Une autre famille de méthodes relève de l'analyse d'impact. Un exemple en est Delphi (Linstone et Turoff, 1975). L'analyse des scénarii est une démarche qui permet de mesurer l'impact de l'interaction de technologies composant un portefeuille de technologies. C'est le cas du modèle Scenario-Based Assessment Model (SBAM) (Banuls et Salmeron, 2007). L'évaluation des risques liés à la technologie est une approche qui tend à mesurer les « synergies négatives », et qui a donné naissance au développement de la méthode ITRACS (Internet-Accessible Technology Risk Assessment Computer System) (Wilhite et Lord, 2006). Mentionnons les méthodes d'analyse de décision dont l'approche la plus représentative est celle proposée par (Saaty 2004) fondée sur l'analyse hiérarchique multicritères.

A la différence des approches décrites plus haut, notre démarche d'évaluation des technologies émergentes intègre trois cadres conceptuels : les systèmes d'information complexes, la théorie des systèmes et la norme ISO 25000 (SQuaRE)[2] pour la qualité des logiciels.

3. Notre approche

Notre objectif est de définir une approche de guidage pour l'évaluation d'un système d'information à base de technologies émergentes. Dans une première section, nous présentons la hiérarchie multicritères que nous avons définie pour organiser l'évaluation. Dans la deuxième partie, nous décrivons l'approche proposée.

3.1. Une hiérarchie multicritères pour l'évaluation d'une technologie émergente

Appréhender une technologie émergente comme un système d'information complexe requiert l'analyse des différentes caractéristiques de ce système. De nombreuses définitions d'un système ont été proposées. Dans une première étape de notre approche, nous nous sommes concentrés sur la forme canonique proposée par Jean-Louis Le Moigne (Le Moigne, 1990). Un système obéit à un but. Il a une structure, qui peut être statique ou dynamique ou qui peut comprendre une partie statique et une partie dynamique. Le système est en interaction avec son environnement. Enfin, un système n'est pas figé : il évolue dans le temps. Ainsi, on peut évaluer une technologie émergente comme un système qui a une structure, un environnement et une évolution. Nous proposons d'organiser notre hiérarchie selon ces trois caractéristiques. Qu'il s'agisse du système lui-même, de son environnement ou de son évolution, de nombreux facteurs entrent en jeu pour le caractériser et pour l'évaluer.

La théorie des systèmes d'information complexes s'appuie sur la perspective sociotechnique des systèmes d'information qui permet de distinguer les facteurs sociaux des facteurs techniques. L'adjectif social est à prendre ici au sens large. Il

[2] http://iso25000.com/index.php/en/iso-25000-standards

englobe tant la dimension organisationnelle que la dimension humaine ou encore la dimension économique et financière. De même le facteur technique couvre tous les aspects de la technologie émergente, tant matérielle que logicielle par exemple. Ainsi notre deuxième niveau d'organisation de la hiérarchie consiste à appréhender le système, son environnement et son évolution d'une part sur le plan social et, d'autre part, sur le plan technologique.

L'organisme de normalisation ISO a élaboré une norme appelée SQuaRE (Software QUAlity Requirements and Evaluation) pour l'évaluation des logiciels. Cette norme s'appuie sur un modèle de qualité en huit dimensions qui sont principalement techniques (six) et fonctionnelles (deux). Issues du modèle de Mc Call (Mc Call, 2002), elles matérialisent les trois types de facteurs (fonctionnement, évolutivité, maintenabilité) préconisés par ce modèle. De cette façon, elles sont aussi en alignement avec les dimensions préalablement considérées pour la description du système d'information complexe.

Ainsi, en considérant successivement, la technologie émergente, comme un système, puis comme un système sociotechnique, puis comme un logiciel, on obtient une hiérarchie en trois niveaux principaux qui peuvent ensuite être affinés (Figure 1). Les huit dimensions de la norme SQuaRE (pertinence fonctionnelle, utilisabilité, fiabilité, sécurité, portabilité, maintenabilité, performance, compatibilité) sont ensuite subdivisées en une trentaine de sous-caractéristiques qui ont été intégrées à la hiérarchie.

Par un processus de mise en correspondance (« mapping ») et de fusion (« merging »), dans un but de complétude, on a ensuite aligné la hiérarchie avec notamment celles proposées dans (Wu *et al.*, 2011) et dans (Huang et Yuan, 2011). Ce processus nous a permis de :

1) Prendre en compte l'environnement économique du système avec les caractéristiques de risque financier, rentabilité, coût total de possession, valeur ajoutée, part de marché.

2) Intégrer l'aspect réglementaire avec la conformité à la loi, à la réglementation du secteur et, le cas échéant, aux certifications en vigueur.

3) Enrichir la hiérarchie avec l'aspect sociétal dans ses dimensions éthique et environnementale.

Sans prétendre à l'exhaustivité, nous présentons le modèle résultant à la figure 1.

Le système	Social	Pertinence fonctionnelle	Complétude fonctionnelle	
			Exactitude	
			Adéquation fonctionnelle	
	Technique	Fiabilité	Maturité	
			Disponibilité	
			Tolérance aux fautes	
			Recouvrabilité	
		Sécurité	Confidentialité	
			Intégrité	
			Non-répudiation	
			Imputabilité	
			Authenticité	
		Performance	Efficience temporelle	
			Utilisation des ressources	
			Capacité	
Son environnement	Social	Humain *Utilisabilité*	Adéquation aux utilisateurs	
			Reconnaissabilité	
			Facilité d'apprentissage	
			Opérabilité	
			Protection des erreurs	
			Esthétique de l'interface	
			Accessibilité	
		Organisationnel	Maîtrise des compétences techniques	
		Economique	Risque financier	
			Rentabilité	
			Coût total de possession	
			Valeur ajoutée	
			Part de marché	
		Sociétal	Acceptabilité éthique	
			Acceptabilité environmentale	
		Réglementaire	Conformité à la loi	Respect de la vie privée
				Droit de la propriété intellectuelle
			Conformité à la réglementation du secteur	
			Respect des certifications	
	Technique *Compatibilité*	Coexistence		
		Interopérabilité		
Son évolution	Sociale	Adaptabilité organisationnelle		
		Adaptabilité fonctionnelle		
		Adaptabilité réglementaire		
		Adaptabilité sociétale		
	Technique	Portabilité	Adaptabilité	
			Installabilité	
			Remplaçabilité	
		Maintenabilité	Modularité	
			Réutilisabilité	
			Analysabilité	
			Modifiabilité	
			Testabilité	
		Passage à l'échelle		

Figure 1. La hiérarchie multicritères d'évaluation

3.2. La méthode de guidage

Face à une technologie émergente, le décideur doit trouver l'information pertinente pour comprendre les enjeux, les composants, les opportunités mais aussi les risques associés. Il lui faut ensuite organiser cette information pour comprendre et, le cas échéant, se faire aider par des experts pour l'évaluation. Il peut ensuite synthétiser cette information.

Le processus proposé comporte ainsi quatre étapes décrites ci-dessous (Figure 2).

Figure 2. Processus d'évaluation

3.2.1. Alimentation de la hiérarchie d'évaluation

Par un « parsing » pour le moment manuel, l'alimentation consiste à engranger la documentation sur la technologie émergente, qu'il s'agisse de presse professionnelle, de livres blancs, d'articles de recherche techniques ou organisationnels. Toutes les données et informations recueillies à partir de ces sources et jugées pertinentes sont transférées dans les nœuds de la hiérarchie. Ce processus est conduit jusqu'à saturation, c'est-à-dire tant qu'il reste de la documentation non parcourue et/ou que des éléments nouveaux sont encore découverts. C'est une phase qui a pour vocation de rassembler et structurer l'information d'aide à la décision.

3.2.2. Evaluation à l'aide de la hiérarchie

Selon son niveau d'expertise, à l'aide de la hiérarchie, l'utilisateur parcourt cette dernière pour :

– porter un jugement sur chaque aspect renseigné,

– détailler et enrichir la hiérarchie pour les aspects jugés plus pertinents,

– compléter, le cas échéant, les éléments manquants, si nécessaire par interrogations d'experts.

3.2.3. Edition du rapport d'évaluation

Après l'élagage et le raffinement de l'arborescence, un rapport ainsi structuré peut être édité.

3.2.4. Capitalisation de la connaissance

La hiérarchie elle-même peut être enrichie des nouvelles branches obtenues lors de son utilisation, permettant ainsi une capitalisation des nouveaux facteurs d'évaluation proposés par les experts.

Le modèle et la méthode ont été utilisés pour évaluer la blockchain, en vue de montrer la faisabilité et l'utilité de l'approche. Cette illustration est expliquée dans la section suivante.

4. Application à la blockchain

Créés en 2008 par Satoshi Nakamoto, le concept et la technologie de blockchain (BC) est le résultat de la combinaison des mécanismes de cryptographie et de l'architecture pair à pair (P2P). La blockchain est considérée comme une innovation disruptive qui a le potentiel de redéfinir de nombreux secteurs de l'économie, des marchés financiers, des activités publiques et gouvernementales, ainsi que des entreprises de haute technologie. A l'origine, cette combinaison était consacrée au développement du « bitcoin », considérée comme une monnaie virtuelle.

La technologie qui sous-tend le bitcoin s'appelle la blockchain. Cette dernière fonctionne comme une base de données publique ou un grand livre de compte ouvert où sont enregistrés les détails de chaque échange de bitcoins. Cette technologie est conçue de manière à empêcher que le même bitcoin ne soit comptabilisé en double, et ce sans qu'aucun intermédiaire, une banque par exemple, n'intervienne. La blockchain enregistre notamment un ensemble de données comme une date et une signature cryptographique associée à l'expéditeur. Dans le cas du bitcoin, il s'agit du nombre de bitcoins envoyés, mais ce pourrait être l'empreinte cryptographique numérique, appelée « fonction de hachage », de n'importe quel document électronique.

Le principe de la blockchain réside dans le fait que chaque opération se trouve inscrite dans des milliers de Grands Livres de compte, chacun soumis à la scrutation d'un observateur différent. Toute blockchain est un registre (et donc un fichier) existant en de très nombreux exemplaires. Les deux paramètres principaux sont la longueur de la blockchain et le nombre d'exemplaires. Pour le bitcoin, la longueur de la blockchain est passée de 27 Go début 2015 à 74 à la mi-2016. On définit le concept de la blockchain comme un système permettant d'enregistrer des transactions. Ce système est fiable puisqu'il est fondé sur la cryptographie. Il est aussi résilient grâce à l'architecture P2P.

À partir de 2014, le concept de blockchain est étendu à des secteurs nécessitant d'enregistrer des transactions ou des contrats (Tsai *et al.*, 2016 ; Yli-Huumo *et al.*, 2016) . A titre d'exemple, mentionnons le site web Proof of Existence[3] qui permet à un utilisateur de télécharger n'importe quel document et d'enregistrer son empreinte pour toujours dans la blockchain. Cette opération permet de prouver que la personne qui a téléchargé le document avait ce document précis en sa possession à un moment donné. Cela peut aussi être utilisé pour prouver que le document n'a pas été modifié depuis ce moment. La startup Stamperya a transformé ce service en une entreprise commerciale qui permet aux autres entreprises « d'affranchir numériquement » n'importe quel document électronique ou e-mail de façon à en établir la propriété et

[3] https://proofofexistence.com/

l'intégrité. L'organisme fédéral américain de réglementation et de contrôle des marchés financiers, le Securities and Exchange Commission, a approuvé l'utilisation de la blockchain comme registre de propriété d'actions par le site de e-commerce Overstock.com. Ce dernier compte utiliser le système technologique de trading alternatif proposé par To.com pour permettre aux particuliers d'acheter et de vendre des actions. L'attrait de ce système tient au fait qu'il offre un règlement immédiat alors que les sociétés de bourse traditionnelles proposent un délai de règlement de 3 jours. Mentionnons aussi la startup Slock.it dont l'idée est d'intégrer le mécanisme de la blockchain via la chaine Ethereum dans des appareils physiques. Il existe aussi des applications de type « smart contracts » reposant sur l'Internet des objets. Un nombre d'objets connectés très élevé est envisagé, probablement 30 milliards au total en 2020 (McKinsey). Notons que le 16 février 2016, le NASDAQ a lancé un projet visant à enregistrer le vote des actionnaires sur la bourse de Tallin en utilisant la blockchain. Enfin, le 1er mai 2016, l'État du Delaware a annoncé que la blockchain doit remplacer les écritures de cet Etat.

4.1. Alimentation de la hiérarchie d'évaluation

Le « parsing » des documents relatifs à la blockchain conduit à une sélection et annotation des paragraphes ou phrases ou expressions, annotation qui les rapproche d'un nœud de la hiérarchie. A titre d'illustration, une partie du texte précédent est représenté ci-dessous après parsing. Les annotations sont en italiques. Les phrases non pertinentes sont barrées.

Créés en 2008 par Satoshi Nakamoto, le concept et la technologie de blockchain (BC) est le résultat de la combinaison des mécanismes de cryptographie et de l'architecture pair à pair (P2P) *<Le système/Technique>*. La blockchain est considérée comme une innovation disruptive qui a le potentiel de redéfinir de nombreux secteurs de l'économie, des marchés financiers, des activités publiques et gouvernementales, ainsi que des entreprises de haute technologie *<Le système/Social>*. ~~A l'origine cette combinaison était consacrée au développement du bitcoin, considérée comme une monnaie virtuelle~~.

~~La technologie qui sous-tend le bitcoin s'appelle la blockchain~~. Cette dernière fonctionne comme une base de données publique ou un grand livre de compte ouvert où sont enregistrés les détails de chaque échange de bitcoins *<Le système/Social>*. Cette technologie est conçue de manière à empêcher que le même bitcoin ne soit comptabilisé en double, et ce sans qu'aucun intermédiaire, une banque par exemple, n'intervienne *<Le système/Technique/Sécurité/Intégrité>*. La blockchain enregistre notamment un ensemble de données comme une date, et une signature cryptographique associée à l'expéditeur. Dans le cas du bitcoin, il s'agit du nombre de bitcoins envoyés, mais ce pourrait être l'empreinte cryptographique numérique, appelée « fonction de hachage », de n'importe quel document électronique *<Le système/Social/ Pertinence fonctionnelle>*.

Le principe de la blockchain réside dans le fait que chaque opération se trouve inscrite dans des milliers de Grands Livres de compte, chacun soumis à la scrutation d'un observateur différent. Toute blockchain est un registre (et donc un fichier) existant en de très nombreux exemplaires *<Le système/Social/Pertinence fonctionnelle>*. Les deux paramètres principaux sont la longueur de la blockchain et le nombre d'exemplaires. Pour le bitcoin, la longueur de la blockchain est passée de 27 Go au début de 2015 à 74 à la mi 2016 *<Le système/Technique/Performance/Capacité>*. On définit le concept de la blockchain comme un système permettant d'enregistrer des transactions *<Le système/Social/Pertinence fonctionnelle>*. Ce système est fiable puisqu'il est fondé sur la cryptographie *<Le système/Technique/Fiabilité>*. Il est aussi résilient grâce à l'architecture P2P *<Le système/Technique/Sécurité/ Non-répudiation>.*

À partir de 2014, le concept de blockchain est étendu à des secteurs nécessitant d'enregistrer des transactions ou des contrats. A titre d'exemple, mentionnons le site web Proof of Existence qui permet à un utilisateur de télécharger n'importe quel document et d'enregistrer son empreinte pour toujours dans la blockchain. Cette opération permet de prouver que la personne qui a téléchargé le document avait ce document précis en sa possession à un moment donné. Cela peut aussi être utilisé pour prouver que le document n'a pas été modifié depuis ce moment *<Le système/Social/Pertinence fonctionnelle>*. La startup Stamperya a transformé ce service en une entreprise commerciale qui permet aux autres entreprises « d'affranchir numériquement » n'importe quel document électronique ou e-mail de façon à en établir la propriété *<Le système/Technique /Sécurité/Imputabilité>* et l'intégrité *<Le système/Technique/ Sécurité/Intégrité>*. L'organisme fédéral américain de réglementation et de contrôle des marchés financiers, le Securities and Exchange Commission, a approuvé l'utilisation de la blockchain comme registre de propriété d'actions par le site de e-commerce Overstock.com. Ce dernier compte utiliser le système technologique de trading alternatif proposé par To.com pour permettre aux particuliers d'acheter et de vendre des actions *<Son environnement/Social/ Réglementaire>*. L'attrait de ce système tient au fait qu'il offre un règlement immédiat alors que les sociétés de bourse traditionnelles proposent un délai de règlement de 3 jours *<Son environnement/ Social/Economique/Rentabilité>….*

Le résultat de ce « parsing » est reporté sur les figures 2 et 3. Même si le processus de « parsing » est, pour le moment, manuel, son application nous a paru aisée, permettant ainsi l'engrangement rapide d'une documentation abondante. La hiérarchie facilite la structuration et permet une détection immédiate de la saturation de la hiérarchie. Ainsi, que la documentation disponible sur la technologie émergente soit abondante ou non, la méthode est applicable.

4.2. Evaluation à l'aide de la hiérarchie

La deuxième étape consiste à analyser la hiérarchie ainsi renseignée et à porter un jugement sur chaque nœud.

<table>
<tr><td rowspan="13">Le système</td><td rowspan="4">Social :
La blockchain est considérée comme une innovation disruptive qui a le potentiel de redéfinir de nombreux secteurs de l'économie, des marchés financiers, des activités publiques et gouvernementales, ainsi que des entreprises de haute technologie. Cette dernière fonctionne comme une base de données publique ou un grand livre de compte ouvert où sont enregistrés les détails de chaque échange de bitcoins</td><td rowspan="4">Pertinence fonctionnelle :
La blockchain enregistre notamment un ensemble de données comme une date, et une signature cryptographique associée à l'expéditeur. Dans le cas du bitcoin, il s'agit du nombre de bitcoins envoyés, mais ce pourrait être l'empreinte cryptographique numérique, appelée « fonction de hachage », de n'importe quel document électronique.
Le principe de la blockchain réside dans le fait que chaque opération se trouve inscrite dans des milliers de Grands Livres de compte, chacun soumis à la scrutation d'un observateur différent. Toute blockchain est un registre (et donc un fichier) existant en de très nombreux exemplaires
On définit le concept de la blockchain comme un système permettant d'enregistrer des transactions
À partir de 2014, le concept de blockchain est étendu à des secteurs nécessitant d'enregistrer des transactions ou des contrats. A titre d'exemple, mentionnons le site web Proof of Existence qui permet à un utilisateur de télécharger n'importe quel document et d'enregistrer son empreinte pour toujours dans la blockchain bitcoin. Cette opération permet de prouver que la personne qui a téléchargé le document avait ce document précis en sa possession à un moment donné. Cela peut aussi être utilisé pour prouver que le document n'a pas été modifié depuis ce moment</td><td>Complétude fonctionnelle</td></tr>
<tr><td>Exactitude</td></tr>
<tr><td>Adéquation fonctionnelle</td></tr>
<tr><td></td></tr>
<tr><td rowspan="9">Technique :
Créés en 2008 par Satoshi Nakamoto, le concept et la technologie de blockchain (BC) est le résultat de la combinaison des mécanismes de cryptographie et de l'architecture pair à pair (P2P)</td><td rowspan="4">Fiabilité :
Ce système est fiable puisqu'il est fondé sur la cryptographie</td><td>Maturité</td></tr>
<tr><td>Disponibilité</td></tr>
<tr><td>Tolérance aux fautes</td></tr>
<tr><td>Recouvrabilité</td></tr>
<tr><td rowspan="5">Sécurité</td><td>Confidentialité</td></tr>
<tr><td>Intégrité :
Cette technologie est conçue de manière à empêcher que le même bitcoin ne soit comptabilisé en double, et ce sans qu'aucun intermédiaire, une banque par exemple, n'intervienne
La startup Stamperya a transformé ce service en une entreprise commerciale qui permet aux autres entreprises « d'affranchir numériquement » n'importe quel document électronique ou e-mail de façon à en établir l'intégrité</td></tr>
<tr><td>Non-répudiation :
Il est aussi résilient grâce à l'architecture P2P</td></tr>
<tr><td>Imputabilité :
La startup Stamperya a transformé ce service en une entreprise commerciale qui permet aux autres entreprises « d'affranchir numériquement » n'importe quel document électronique ou e-mail de façon à en établir la propriété</td></tr>
<tr><td>Authenticité</td></tr>
<tr><td rowspan="3">Performance</td><td>Efficience temporelle</td></tr>
<tr><td>Utilisation des ressources</td></tr>
<tr><td>Capacité :
Les deux paramètres principaux sont la longueur de la blockchain et le nombre d'exemplaires. Pour le bitcoin, la longueur de la blockchain est passée de 27 Go au début de 2015 à 74 à la mi 2016</td></tr>
</table>

Figure 2. Evaluation de la blockchain (système)

Cinq types de jugements peuvent être émis : 1) le nœud ne contient pas d'information : cette situation peut traduire un manque d'information sur ce critère ou le caractère non pertinent du critère pour cette technologie, 2) le nœud contient une information factuelle (grisé) à valeur descriptive, 3) le nœud traduit un jugement positif, une opportunité apportée par la technologie (vert), 4) le nœud représente une alerte (orange), informant le décideur sur un aspect nécessitant une surveillance particulière, 5) le nœud représente un risque (rouge) qui appelle à une évaluation renforcée.

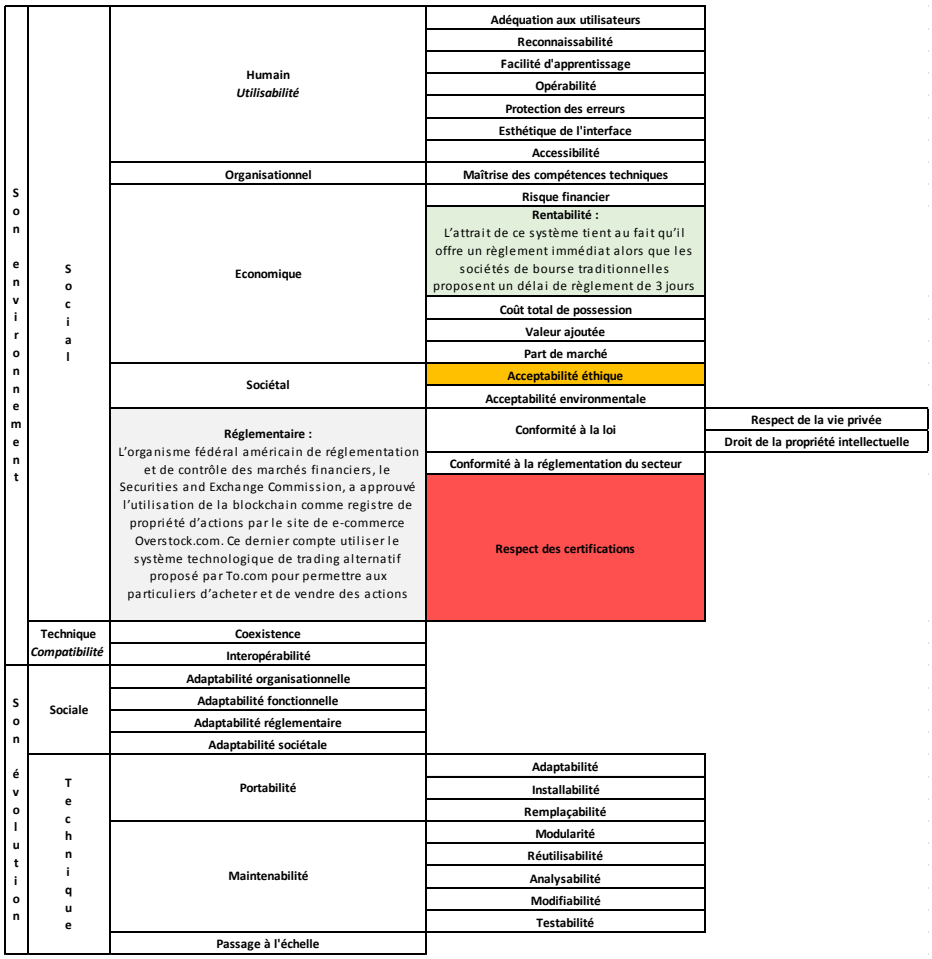


Figure 3. Evaluation de la blockchain (environnement et évolution)

L'évaluation ainsi menée de la blockchain fait ressortir un manque de documentation sur de nombreux aspects. Un effort de documentation plus complète

permettrait de disposer de plus d'éléments d'évaluation. Toutefois, on constate que plusieurs traits saillants de la blockchain sont mis en évidence notamment les nombreuses opportunités qu'elle représente (opportunité sociale, fiabilité, intégrité, non-répudiation, rentabilité) mais aussi les risques (maturité, respect des certifications). Ainsi, un des freins actuels dans son développement est celui relatif aux certifications que les entreprises, et notamment les banques, doivent respecter.

La confrontation de la documentation avec la hiérarchie permet de faciliter l'évaluation de la technologie. L'utilisateur non expert peut évaluer chaque information renseignée à l'aide de l'échelle de valeurs décrite plus haut. L'expert peut, quant à lui, valider cette évaluation, détailler certains nœuds qu'il juge trop synthétiques et ajouter des évaluations aux nœuds non renseignés, le cas échéant.

En l'état, la hiérarchie ainsi complétée sert de base à l'édition du rapport d'évaluation. Enfin, dans le cas où l'expert a complété la hiérarchie, on peut entériner ou non les nouvelles branches (phase de capitalisation de la connaissance). Dans l'exemple de la blockchain, l'acceptabilité éthique peut être affinée par l'expert et conduire à un enrichissement de la hiérarchie utile pour d'autres technologies émergentes.

5. Conclusion et recherches futures

Nous avons présenté dans cet article une démarche d'évaluation des systèmes d'information à base de technologies émergentes. Cette démarche, fondée sur le constat que les technologies émergentes sont des systèmes complexes, combine trois cadres conceptuels : la théorie sous-jacente aux systèmes d'information complexes, la théorie des systèmes et la norme ISO 25000 dédiée à la qualité des logiciels. Afin de structurer l'évaluation, nous avons développé une hiérarchie multicritères. Nous avons tenu compte des dimensions sociales et techniques pour chaque composante du système à évaluer, notamment le système lui-même, son environnement et son évolution. Ainsi, nous avons retenu des critères tels que la pertinence fonctionnelle pour l'aspect social du système, et la fiabilité, la sécurité et la performance pour l'aspect technique. Pour illustrer cette démarche, nous l'avons appliqué au cas de la «blockchain», qui constitue aujourd'hui une technologie émergente avec de nombreux domaines d'applications.

Nous projetons, en termes de recherche future, d'étendre l'évaluation en associant aux critères des métriques. En effet, chaque critère peut être évalué à l'aide de métriques avec pour objectif d'évaluer la complexité par dimension et de prioriser les actions afin de maîtriser la complexité. A noter que les poids des dimensions d'évaluation et des critères ne sont pas les mêmes selon les secteurs d'activité ou les domaines visés. Ainsi, dans certains cas, l'adaptabilité spatiale (la scalabilité ou passage à l'échelle) peut avoir un poids important. Dans d'autres cas, c'est la conformité réglementaire qui est prépondérante, tout en dépendant des domaines concernés.

Un autre axe de recherche concerne une démarche qui utiliserait le même modèle hiérarchique en rétro-ingénierie, comme trame permettant aux organisations de

déterminer les facteurs les plus importants et les plus en adéquation avec leurs besoins lors du développement de technologies émergentes.

Enfin, une autre voie de recherche consiste à intégrer les techniques d'analyse du langage naturel pour automatiser la phase d'analyse de la documentation.

Bibliographie

Al-Arabiat, D., Ahmad, W. F. W., Sarlan, A. (2015). Review on critical factors of adopting cloud mobile learning. In Technology Management and Emerging Technologies (ISTMET), *2015 International Symposium on (pp. 69-73)*. *IEEE*.

Baccarini, D. The concept of project complexity--a review, *International Journal of Project Management*, vol. 14, issue 4, pp. 201-204, 1996.

Banuls V.A., Salmeron J.L. A Scenario-Based Assessment Model—SBAM, *Technological Forecasting and Social Change*, Volume 74, Issue 6, July 2007, Pages 750–762

Cozzens, S., Gatchair, S., Kang, J., Kim, K. S., Lee, H. J., Ordóñez, G., Porter, A. (2010). Emerging technologies: quantitative identification and measurement. *Technology Analysis & Strategic Management*, 22(3), 361-376.

Day, G. S., Schoemaker, P. J., Gunther, R. E. (2004). *Wharton on managing emerging technologies*. John Wiley & Sons.

Emmeche, C., Koppe, S. Stjernfelt (1997) F. Explaining Emergence: Towards an Ontology of Levels. *Journal for General Philosophy of Science* 28: 83.

Fitsilis P., Measuring the complexity of software projects, *World Congress on Computer Science and Information Engineering*, 2009.

Geraldi, J., Adlbrecht, G. (2008). On faith, fact, and interaction in projects. *IEEE Engineering Management Review*, 2(36), 35-49.

Holland, J. H. (2006). Studying complex adaptive systems. *Journal of Systems Science and Complexity*, 19(1), 1-8.

Huang L., Yuan Y. (2010). Evaluation on the industrialization potential of emerging technologies based on principal component and cluster analysis. In : *Computer Modelling and Simulation (UKSim), 2010 12th International Conference on. IEEE*, p. 317-322.

Kim, R. M., Kaplan, S. M. (2006). Interpreting socio-technical co-evolution: Applying complex adaptive systems to IS engagement. *Information Technology & People*, 19(1), 35-54.

Le Moigne, J. L. (1990). *La modélisation des systèmes complexes*. Paris: Bordas, Dunod, 1990.

Letaba, P. T., Pretorius, M. W., Pretorius, L. (2014). The use of bibliometrics in the development of technology roadmaps: Planning for industrial impact of emerging technologies. *In Engineering, Technology and Innovation (ICE), 2014 International ICE Conference on* (pp. 1-8). IEEE.

Linstone, H. A., Turoff, M. (Eds.). (1975) The Delphi method: Techniques and applications (Vol. 29). Reading, MA: Addison-Wesley.

Lloyd S. Measures of complexity: a non exhaustive list. *Control Systems Magazine*, IEEE, 21:7–8, 2001.

Mala M., Çil I. A Taxonomy for Measuring Complexity in Agent-Based Systems, 2011 IEEE *2nd International Conference on Software Engineering and Service Science*, Pages: 851 – 854.

McCall, J. A. 2002. Quality Factors. *Encyclopedia of Software Engineering*.

Owen, C. L. (2007). Evaluation of complex systems. *Design Studies*, 28(1), 73-101.

Perminova, O., Gustafsson, M., Wikström, K. (2008). Defining uncertainty in projects–a new perspective. *International Journal of Project Management*, 26(1), 73-79.

Rotolo, D., Hicks, D., & Martin, B. R. (2015). What is an emerging technology?. *Research Policy*, 44(10), 1827-1843.

Saaty, T. L. (2004). Decision making—the analytic hierarchy and network processes (AHP/ANP). *Journal of systems science and systems engineering*, 13(1), 1-35.

Schlauderer S., Overhage S., Weidinger J. (2016). New Vistas for Firefighter Information Systems? Towards a Systematic Evaluation of Emerging Technologies from a Task-Technology Fit Perspective. In *System Sciences (HICSS), 2016 49th Hawaii International Conference on (pp. 178-187). IEEE*.

Skyttner, L. *General Systems Theory: Problems, Perspectives, Practice*. 2nd ed. Singapore: World Scientific, 2005.

Sommerville I., Cliff D., Calinescu R., Keen J., Kelly T., Kwiatkowska M., Paige R. (2012). Large-scale complex IT systems. *Communications of the ACM*, 55(7), 71-77.

Song Y., Yin L., Research on species traits of emerging technologies and the path of formation, *Management*, 2007, 4(2), pp. 211-215.

Tran T.A, Daim T. A taxonomic review of methods and tools applied in technology assessment, *Technological Forecasting & Social Change* 75 (2008) 139–1405

Tsai, W. T., Blower, R., Zhu, Y., Yu, L. (2016). A System View of Financial Blockchains. In *Service-Oriented System Engineering* (SOSE), 2016 IEEE Symposium on (pp. 450-457).

Wang K., Yu J., Yu Y., Qian Y., Zeng D., Guo S., Xiang Y., Wu J., "A survey on energy internet: Architecture, approach, and emerging technologies," *IEEE Systems Journal*, no. 99, vol. PP, pp. 1–14, 2017.

Wilhite, A., Lord, R. (2006). Estimating the risk of technology development. Engineering *Management Journal*, 18(3), 3-10.

Wu F., Feng H., Huang L. (2011). Reviews on economic effects assessment of emerging technologies. In *IT in Medicine and Education (ITME), 2011 International Symposium on* (Vol. 2, pp. 304-308). IEEE.

Yli-Huumo, J., Ko, D., Choi, S., Park, S., Smolander K. (2016). Where Is Current Research on Blockchain Technology? — A Systematic Review. *PloS one*, 11(10), e0163477.

Processus de conduite de la recherche et ingénierie des processus : vers une fertilisation croisée

Nadine Mandran, Sophie Dupuy-Chessa, Eric Céret

Université de Grenoble Alpes, Laboratoire d'informatique de Grenoble, CNRS
F-38000 Grenoble, France
Prenom.Nom@univ-grenoble-alpes.fr

RESUME.
La conduite de la recherche en informatique centrée humain nécessite un guidage pour les chercheurs dans l'élaboration et l'évaluation de leur contribution. Pour répondre à ce besoin, nous proposons le processus de conduite de la recherche THEDRE. Afin d'évaluer THEDRE, nous utilisons Promote, une taxonomie des processus de développement logiciel. Ce travail nous a permis non seulement d'identifier certaines limites de THEDRE, mais aussi celles de Promote. Cette fertilisation croisée entre ingénierie des processus et conduite de la recherche nous permet de poser les bases d'une taxonomie spécifique aux processus de conduite de la recherche centrés humain.

ABSTRACT.
Leading research in human centred computer science needs guidance for helping researchers in elaborating and evaluating their contribution. With this goal, we propose THEDRE, a process to lead research in human centred computer science. In order to evaluate THEDRE, we use Promote, a taxonomy for Software Development Process Model. This work allows us to identify limits of THEDRE, but also of Promote. This cross fertilization between process engineering and research leading permits to identify the foundation of a taxonomy specific to human centred research processes.

MOTS-CLES : processus, conduite de la recherche, taxonomie, modèle de concepts

KEYWORDS: process, research methodology, taxonomy, concept model

1. Introduction

La conduite de la recherche en informatique nécessite de faire appel à des parties prenantes pour construire et évaluer une connaissance scientifique puisque les systèmes et logiciels conçus s'adressent in fine à des utilisateurs. Ce type de recherche est donc confronté à l'intégration de l'humain et de son environnement familial, professionnel, etc. Nous la nommerons Recherche Informatique Centrée Humain (RICH) dans le sens où l'utilisateur est central pour la construction et

l'évaluation de la connaissance scientifique en Informatique. Ce centrage rend complexe la RICH car elle nécessite d'utiliser des démarches expérimentales inspirées des sciences humaines et sociales auxquelles ne sont pas formés les chercheurs en informatique.

La complexité en RICH incite à proposer des méthodes de conduite de recherche qui décrivent un processus répétable et adaptable pour offrir du guidage dans cette démarche difficile. Le processus de conduite de la RICH est particulier dans le sens où il a pour objectif de produire de la connaissance scientifique et un outil qu'un utilisateur peut mettre en œuvre. La construction de la connaissance scientifique et celle de l'outil sont interdépendantes. Un tel processus pour la construction d'une connaissance scientifique, pose également la question de l'ancrage dans un paradigme épistémologique qui définit comment une connaissance est produite et quels sont les critères de validité et de valeur de cette connaissance (Avenier and Thomas, 2015). De plus, un processus de conduite de la RICH demande un certain niveau de traçabilité des activités et des productions pour rendre compte de la validité de la connaissance scientifique. Cette traçabilité est importante car la mesure des représentations de l'humain est instable et inconstante (Jambon, 2009).

Des méthodes de conduite de la recherche (Wang and Hannafin, 2005) (De Vries, 2007) (Hevner, 2007) (Peffers et al., 2006) (Drechsler and Hevner, 2016) ont déjà été proposées en RICH pour répondre à ces problématiques. Mais aucune d'elles n'a été étudiée comme un processus à part entière en mettant en oeuvre les concepts, des techniques et des outils de support proposés en ingénierie des processus.

La méthode THEDRE (Mandran, 2017) – Traceable Human Experiment Design REsearch – est un modèle de processus de RICH, que nous avons créé et que nous souhaitions évaluer. Or nous avions déjà une expérience dans le domaine de l'évaluation des modèles de processus avec Promote (Céret et al., 2013a), une taxonomie des modèles de processus de conception et de développement de systèmes d'information (SI). Dans cet article, nous présentons l'application de Promote – centré sur le développement informatique – au modèle de processus de THEDRE, centré sur la conduite de recherche centrée humain, ainsi que les enseignements que nous avons tirés de cet exercice. Ces enseignements portent à la fois sur les processus de conduite de la recherche, sur Promote, ainsi que sur la fertilisation croisée des deux approches : le méthodologue qui crée des méthodes de conduite de la recherche, peut les analyser et les complèter grâce à l'application des critères de Promote ; pour le chercheur en RICH, l'utilisation de Promote dans un domaine non exploré i.e. celui de la conduite de la recherche enrichit les critères d'analyse des processus.

Dans la suite de cet article, nous définissons les caractéristiques de la RICH qui sont utilisées dans la méthode THEDRE. Nous présentons les axes de PROMOTE pour caractériser un processus. Nous exposons ensuite la méthode THEDRE. Ensuite, nous la mettons en perspective avec le modèle PROMOTE pour élaborer une taxonomie des processus de conduite de la recherche. Avant de conclure, nous discutons des extensions issues de ce travail pour THEDRE et PROMOTE.

2. Caractéristiques d'un processus pour la RICH

La RICH se préoccupe à la fois de produire une connaissance scientifique et des outils activables pour accompagner l'activité humaine (p.ex., un langage, un dictionnaire, une interface, un modèle). Ces outils produits dans la RICH représentent la connaissance scientifique dans une forme utilisable par l'utilisateur. Cette dualité est le propre des sciences de l'artificiel décrite par (H.Simon, 1969). Dans certains cas, l'outil activable est décomposable en sous-parties que nous appelons composants activables. Les différents composants qui composent l'outil activable devront être identifiés pour le construire et l'évaluer lors des phases expérimentales (Gregor and Hevner, 2013),(Mandran et al., 2013).

Les travaux en RICH doivent prendre en compte une dimension pluridisciplinaire et une dimension transversale. Ils sont pluridisciplinaires dans le sens où ils se préoccupent de problématiques en informatique qui doivent mobiliser des humains et donc utiliser des méthodes des sciences humaines et sociales. Ils sont transversaux, car le problème se pose dans différentes spécialités de la recherche en informatique. Ainsi, nous avons pu observer le problème dans 4 spécialités: Interaction Homme-Machine (IHM), Environnements Informatiques pour l'Apprentissage Humain (EAIH), Systèmes d'Information (SI), et robotique. Ces deux caractéristiques, pluridisciplinarité et transversalité, sont les premiers éléments de complexité du problème (Jean-Daubias, 2004) et nous conduisent à préciser la vision du processus telle qu'elle est déployée dans THEDRE. Nous traitons en particulier les caractéristiques suivantes :

- Des **processus itératifs** pour construire l'outil : l'outil ici étant informatique, il est soumis, pour sa réalisation, aux préconisations des méthodes de conception et de développement, qui, le plus souvent, sont itératives, comme les méthodes agiles (Martin, 2003).
- **La complexité du terrain à investiguer** est liée à l'une des spécificités en RICH qui est de construire et d'évaluer un instrument avec des utilisateurs. Cette caractéristique impacte le processus scientifique au point que (Sein et al., 2011) constatent qu' « une nouvelle méthode recherche est nécessaire, qui reconnaisse que l'artefact émerge de l'interaction avec le contexte organisationnel même quand sa conception initiale était guidée par les intentions du chercheur ».
- **La combinaison des méthodes de production et d'analyse de données** : les méthodes quantitatives/statistiques s'appliquent principalement en phase d'évaluation mais ne sont pas adéquates lors de la construction de l'instrument. Il est alors nécessaire d'envisager des méthodes d'investigation pour recueillir l'avis des utilisateurs. Il faut donc utiliser des méthodes qualitatives pour comprendre, explorer (Paille and Mucchielli, 2011) et des méthodes quantitatives pour quantifier, valider (Howell et al., 2007).

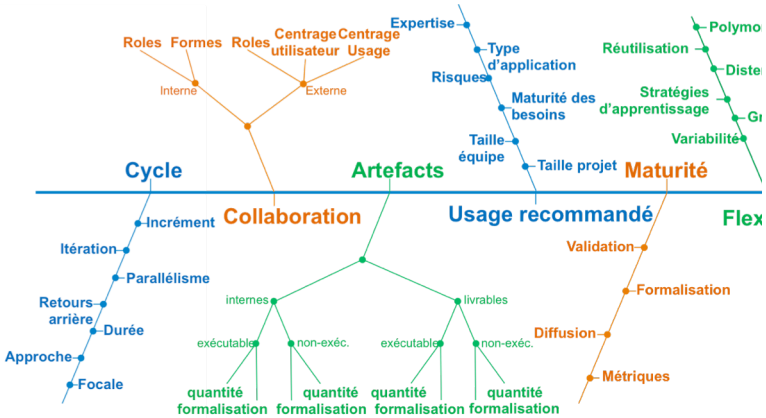
L'objectif dual de la RICH et ces trois points interrogent directement la manière de construire la connaissance scientifique autrement dit de se positionner dans un paradigme épistémologique pour expliciter les hypothèses de construction de cette

connaissance et les critères de valeurs et de validité de la connaissance. Nous nous situons dans le constructivisme pragmatique (Avenier and Thomas, 2015) qui définit la connaissance produite comme un raffinement des connaissances existantes et avec une prise en compte d'un contexte d'application. La validité et la valeur de la connaissance repose sur la multiplicité des données produites à l'aide du terrain.

3. Taxonomie des processus : Promote

En considérant que les processus de conduite de la recherche ont des caractéristiques similaires aux processus de conception de logiciels, nous allons utiliser les caractéristiques de Promote (Process Models Taxonomy for Enlightening choices) (Céret et al., 2013a) pour identifier les points importants pour un modèle de processus de conduite de RICH. Promote est une taxonomie de modèles de processus de conception et développement logiciel. Elle a été créée afin de pouvoir catégoriser et comparer ces modèles. Elle permet aussi d'identifier des points d'amélioration pour permettre ainsi leur enrichissement. Elle a été mise en œuvre en ce sens à plusieurs reprises (Céret et al., 2013a), (Céret et al., 2013b), (Céret, 2014).

Promote comporte six axes principaux pour caractériser un modèle de processus (Figure 1Figure 1). Caractériser un modèle de processus de développement (SDPM – Software Development Process Model) se fait sur la base de ce qui est défini dans le processus et non des habitudes de mise en œuvre. Chacune des caractéristiques (les feuilles de l'arête dans la figure ci-dessus) est associée à une graduation qui permet d'exprimer à quel point un SDPM met en œuvre la



caractéristique analysée.

Figure 1 : Promote caractérise les processus selon 6 axes principaux divisés en sous-axes

L'axe du **cycle** de vie décrit l'organisation interne du cycle proposé par le SDPM et comporte sept sous-axes qui caractérisent les aspects *itératifs* et *incrémentaux* du processus. L'axe du cycle définit aussi la capacité de ce dernier à supporter des tâches menées à en *parallèle*, les éventuelles possibilités de *retour en arrière* (et les procédures de gestion associées), la *durée* du processus si elle est définie, son *type d'approche* (ascendante, descendante, composite) et enfin sa *focale* c'est-à-dire le ou les principaux aspects selon lesquels le processus est décrit : les activités, les produits, les stratégies, les buts ou les décisions (Rolland, 2005).

La **collaboration** analyse comment les différentes parties prenantes sont supposées travailler ensemble. Au niveau interne, c'est-à-dire entre les membres de l'équipe de conception et de développement, Promote suggère de mesurer le *nombre de rôles*, ainsi que la ou les *formes de collaboration* recommandées. Pour les autres parties prenantes (niveau "externe"), Promote propose aussi de quantifier les *rôles*, et de mesurer la proportion d'activités *centrées sur l'utilisateur* (et donc à quel point la participation de ces derniers est possible). Enfin la quantité d'activités *centrées sur l'usage*, c'est-à-dire focalisées sur la modélisation de la représentation visuelle et de l'interaction (Constantine et al., 2003) est évaluée.

Les **artefacts** (ou produits) sont catégorisés comme *internes* (non destinés à être livrés au client) ou, à l'inverse, *livrables*, et subdivisés en produits *documentaires* ou *exécutables*. Pour chacune des quatre catégories, Promote propose de mesurer le niveau de *formalisation* et la *quantité* préconisés par le modèle de processus.

L'**usage recommandé** identifie les contextes pour lesquels le SDPM indique être adapté et repose sur des critères comme la *taille du projet* ciblé (et les indications pour évaluer cette taille), la *taille de l'équipe*, la *maturité des besoins*, le *niveau de risque* compatible avec le processus suggéré, le *type d'application* (ex : systèmes embarqués) et le *niveau d'expertise* attendu de l'équipe.

La **maturité** du modèle de processus est évaluée à travers les validations que ses auteurs indiquent. Promote propose d'étudier le *degré de formalisation* du SDPM, c'est-à-dire la quantité de parties décrites avec des langages formels ou semi-formels. La *diffusion* du SDPM est mesurée à travers le nombre de publications, de sites Internet et de livres qui lui sont consacrés. Enfin, Promote évalue si le SDPM comporte une définition de *métriques* permettant de mesurer sa bonne application, conformément à la préconisation de (Cook and Wolf, 1999).

Le dernier axe de Promote concerne la flexibilité du SDPM, c'est-à-dire sa capacité à proposer des adaptations à la situation locale dans laquelle il est mis en œuvre. La variabilité est la capacité d'un modèle de processus à proposer différents chemins parmi les éléments (activités, produits,...) qui le composent. La granularité mesure la capacité du modèle de processus à être instancié avec différentes granularités (tâches, sous-tâches,…). Les stratégies d'apprentissage identifient ce que le SDPM comporte pour faciliter sa prise en main par un novice (exemples d'application, exercices,…). La distensibilité évalue la capacité du modèle de

processus a être étendu (p.ex. procédure pour ajouter une activité,…). La réutilisation étudie si le modèle de processus fournit des éléments sur lesquels l'équipe peut s'appuyer (p.ex. des patrons de conception, librairies,…). Enfin, le polymorphisme d'un SDPM est sa capacité à être présenté selon différentes focales, par exemple de pouvoir basculer d'une perspective centrée sur les produits à une perspective centrée sur les activités.

4. THEDRE: un processus de conduite de la recherche

Cette section décrit le processus de conduite de la recherche THEDRE. En le considérant comme processus de conception, caractérisé selon les axes de Promote.

4.1. Le cycle de vie de THEDRE

4.1.1 Description du cycle de THEDRE

La vision globale du cycle de THEDRE est donnée par un processus qui suit un cycle d'amélioration Plan-Do-Check-Act (PDCA) (Deming, 1952). Le processus est découpé en cinq sous-processus organisés de manière cyclique et itérative (numérotés de 1 à 5 sur la Figure 2Figure 2). Ces sous-processus sont détaillés par des blocs constitués de tâches que nous ne présenterons pas ici par manque de

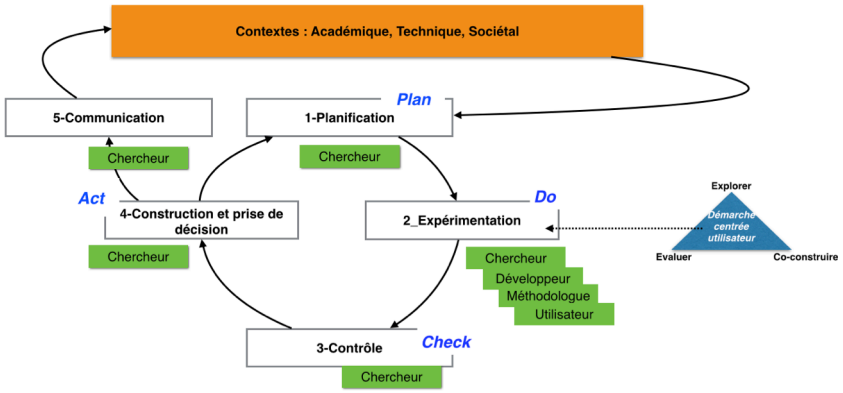


Figure 2 - Vue globale du processus de THEDRE

place.

1 - Planification de la recherche : pour le chercheur, il s'agit ici d'élaborer (en première itération) ou de raffiner (dans les itérations suivantes) la question de recherche et de proposer un outil activable.

2 - Expérimentation : ce sous-processus vise la conception, la mise en œuvre et l'évaluation de l'outil activable. Il nécessite la mise en œuvre de compétences de plusieurs domaines : le chercheur expose sa vision du problème et de l'outil, le

méthodologue propose une démarche expérimentale et le développeur créé l'outil activable à partir des spécifications du chercheur. Enfin, l'utilisateur est intégré dans la phase d'expérimentation selon les trois « actions » de la démarche centrée utilisateur : explorer, co-construire, évaluer (Mandran et al., 2013).

3 - Contrôle : il porte sur la vérification des résultats issus du sous-processus « Expérimentation » pour qu'ils soient acceptables afin de construire la connaissance scientifique et l'outil activable.

4 - Construction et prise de décision : c'est le temps de la construction de la connaissance scientifique, temps où le chercheur confronte ses nouveaux résultats à ceux de la communauté académique et technique. Le chercheur évalue les résultats des expérimentations pour savoir si le travail de recherche et l'outil activable sont suffisamment aboutis et novateurs pour être communiqués ou s'il faut conduire une nouvelle itération des sous-processus 1 à 4.

5 - Communication : elle est relative à la publication des résultats liés à l'instrument de la recherche, dans les contextes académiques, technologiques et sociétales. Ce sous-processus est l'étape finale avant d'itérer sur une nouvelle question de recherche.

4.1.2 Caractérisation du cycle de THEDRE selon Promote

THEDRE a besoin d'être présenté de manière facilement compréhensible pour des chercheurs en RICH. La modélisation des processus sous forme d'activités étant la plus connue, THEDRE privilégie la focale activité1 pour identifier les tâches à conduire avec une identification des produits à réaliser et à tracer.

Le cycle de THEDRE est un processus itératif afin de prendre en compte l'amélioration continue du cycle de Deming. Des itérations sont aussi possibles lors du sous processus « Expérimentation » afin de construire et d'évaluer l'outil activable. On peut donc dire que le processus de THEDRE suggère des itérations globales (sur l'ensemble du cycle) et régionales (au niveau des sous-processus).

L'incrément est lié à la valeur apportée à la connaissance scientifique et à création de l'outil activable. Si on considère qu'une livraison correspond à la publication d'une connaissance scientifique, celle-ci est le résultat d'une itération globale qui peut être relativement longue, voire qui peut ne pas être satisfaisante et nécessiter une nouvelle itération. On peut donc dire que THEDRE offre un petit nombre d'incréments.

Le parallélisme est défini dans THEDRE entre des blocs de tâches ou entre tâches. Il est clairement spécifié pour les tâches concernées. Par exemple, dans le sous-processus d'expérimentation (que nous n'avons pas détaillé par manque de place), les blocs « Définir les utilisateurs et leur implication dans l'expérience » et « Choisir les méthodes de production de données » peuvent être menés en parallèle.

[1] Les textes en italique sont les graduations de Promote

Le caractère itératif du cycle peut permettre, dans une certaine mesure, les retours arrière, au sens où une nouvelle itération peut permettre une correction. Mais THEDRE ne précise pas de quelle manière prendre en compte ces retours en arrière. Il offre des retours arrière sans procédure spécifiée. De même, la durée d'un cycle n'est pas précisée. Néanmoins, les durées proposées dans le cadre de Promote (de quelques semaines à quelques mois) sont peu pertinentes pour des cycles de recherche qui sont par nature plus longs que les cycles de développement. Ce point serait à revoir dans une taxonomie spécifique aux processus de conduite de la recherche.

Enfin le caractère descendant ou ascendant de l'approche ne semble pas le plus pertinent pour une méthode de conduite de la recherche. En effet, cette question de l'approche interroge, pour un SMDP, l'ensemble des « bagages » avec lesquels le projet est conduit : jusqu'où faudra-t-il décomposer le problème ? De quoi peut-on se servir qui ait déjà été mis en œuvre ? Comment augmenter la probabilité d'aller vers la solution ? Ces questions, dans le domaine de la recherche, relèvent du paradigme épistémologique : il correspond à la manière dont la connaissance scientifique va être construite et évaluée, avec ou sans la prise en compte de l'humain et de son contexte. Le choix d'un tel paradigme par le chercheur justifie la manière dont il va construire et évaluer la connaissance scientifique produite. Ce positionnement doit être défini avant tout démarrage d'un processus de recherche, c'est un des prérequis.

4.2. La collaboration au sein de THEDRE

4.2.1 Description de la collaboration au sein de THEDRE

THEDRE implique différents acteurs qui vont avoir un rôle dans la construction et l'évaluation de l'instrument de la recherche. Ainsi, 4 acteurs interviennent dans le processus de la RICH :

- **Le chercheur en RICH** : il a pour rôle de poser la problématique de recherche à partir de ses connaissances d'un domaine, de faire évoluer la connaissance scientifique et de la communiquer. Il conçoit l'outil activable. Il est garant de la cohérence entre le développeur et le méthodologue.
- **Le développeur** : il a pour rôle de développer l'outil activable quand il nécessite de telles compétences (p.ex., application informatique, site web).
- **Le méthodologue** : il a pour rôle de concevoir, de mettre en œuvre et d'évaluer les expérimentations conduites avec les utilisateurs. C'est l'acteur qui apporte la compétence en méthodes de production et d'analyse des données pour répondre à une problématique. Il mobilise des travaux des SHS et d'autres domaines tels que les statistiques.
- **L'utilisateur** : il a pour rôle de participer aux expérimentations afin de faire part de ses représentations du « monde connu » afin de construire et d'évaluer l'instrument. Ce terme « utilisateur » doit être compris au sens large, comme utilisateur final d'une application aussi bien que comme décideur. Il intervient dans l'exploration, la co-construction et l'évaluation.

4.2.2 Caractérisation de la collaboration de THEDRE dans Promote

La collaboration interne dans THEDRE est clairement précisée : trois *rôles* sont identifiés (l'utilisateur étant considéré comme un rôle externe) et des activités de collaboration forte (*coopération*) sont proposées. Par exemple, les trois acteurs internes définissent ensemble les tâches relatives à l'expérimentation.

Ces trois types d'acteurs correspondent à un nombre adéquat de rôles d'après Promote (16 rôles ou moins), qui se base sur (Harrison and Coplien, 1996) pour évaluer si le nombre de rôle est favorable ou non à la communication interne. Cet indicateur n'est pas spécifique à une équipe de développement et reste donc à priori valide pour une équipe de recherche.

En matière de collaboration externe, Promote permet de définir les principales caractéristiques d'un processus de conduite de recherche, même si certains aspects sont à revoir. Un seul *rôle* externe est proposé, celui d'utilisateur, même si, comme nous l'avons vu, ce terme peut recouvrir différents sens. On peut se demander si la spécification d'autres rôles (par exemple, un décideur ou un utilisateur expert), permettrait d'améliorer la prise en compte de l'humain dans la RICH.

Le processus expérimental de THEDRE est *fortement centré utilisateur* : il contient de nombreuses activités avec les utilisateurs, qui correspondent aux différentes étapes d'une démarche centrée utilisateur. Enfin le caractère centré usage d'un processus de conduite de la recherche ne semble pas pertinent ici, car aucun de ces processus ne s'appuie sur des modèles.

4.3. Les artefacts dans THEDRE

4.3.1 Description des artefacts dans THEDRE

Le processus de THEDRE produit des livrables (tableau 1) qui contribuent à la construction et à l'évaluation des composants activables, des outils activables et in fine de la connaissance scientifique. Chaque livrable est lié à un ensemble de tâches cohérentes. Certains des livrables suivent des guides fournis par THEDRE. Par exemple, un protocole expérimental est rédigé en suivant un format prédéfini.

Bloc où est constitué le livrable	No du livrable	Contenu du livrable
Bloc 1, 2 et 9	1	Synthèse sur l'état de l'art, la veille technologique et le contexte sociétal b) description des terrains d'étude et des utilisateurs
Bloc 2	2	Spécification pour le développement de l'outil activable
Bloc 2	3	Base des contacts des utilisateurs
Bloc 2 et 7	4	Liste des indicateurs d'objectifs
Bloc 3	5	Tableau de décomposition de l'outil activable

Tableau 1 : Extrait du tableau des livrables selon les blocs de la méthode THEDRE

4.3.2 Caractérisation des artefacts de THEDRE dans Promote

Dans THEDRE, les artefacts – au sens de Promote – sont des documents, mis à part les outils activables (p.ex., un prototype). La *quantité de livrables exécutable est faible*. Pour les autres artefacts, THEDRE ne préconise pas de nombre spécifique. Cependant, dans le cadre de la traçabilité, les expérimentations et les données associées peuvent être considérées comme des livrables (i.e. elles doivent être accessibles à la communauté scientifique). En ce sens, THEDRE préconise un *grand nombre d'artefacts livrables non-exécutables*. A l'inverse la méthode ne donne pas de préconisation sur les artefacts internes, dans lesquels on pourrait englober tout ce qui a été produit au cours du travail d'idéation sans être remis à la communauté. Enfin THEDRE définit le format de certains artefacts, livrables et indicateurs. On peut donc dire que THEDRE suggère *des artefacts informels et semi-formels*.

4.4. Utilisation recommandée de THEDRE

4.4.1 Utilisation recommandée par THEDRE

THEDRE précise peu de choses sur son utilisation. Le processus s'adresse à des chercheurs, principalement des doctorants ; l'équipe est pluridisciplinaire. Mais aucun élément ne précise le nombre d'intervenants dans ces rôles.

De plus, THEDRE s'adresse clairement à certains champs de la recherche : son utilisation est recommandée dans les domaines qui nécessitent la prise en compte de l'humain, par exemple en interaction homme-machine, systèmes d'information, robotique ou encore en apprentissage humain assisté par ordinateur.

4.4.2 Utilisation recommandée de THEDRE vue par Promote

La plupart des critères de l'axe « Utilisation Recommandée » sont très spécifiques aux SDPM. Ainsi les risques, quelle que soit leur nature (technique, financier, organisationnel,…), ne sont pas ceux relatifs à un travail de recherche, où l'on focaliserait plutôt sur son caractère incertain. De manière similaire, la maturité des besoins n'a pas de sens dans le cadre d'un travail de recherche où il n'y a pas de parties prenantes pour exprimer un besoin.

Promote met cependant en évidence que THEDRE ne précise rien sur la taille des projets concernés, alors que cet élément est important en recherche. La taille du projet pourrait être relative au type de projet (local, ANR, européen,…). Le niveau d'expertise et la taille de l'équipe ne sont pas non plus mentionnés dans THEDRE mais pourraient apporter des informations utiles à sa mise en œuvre.

4.5. Maturité de THEDRE

4.5.1 Description de la maturité de THEDRE

THEDRE est issu d'une dizaine d'années de travail en accompagnement de la recherche. Sa valeur et sa validité reposent sur de nombreuses mises à l'épreuve : 25

travaux de thèses dans 4 domaines (SI, IHM, EIAH, robotique). THEDRE a été construit avec une méthode d'observation participante. Ces différents travaux ont donné lieu à plus de 15 des publications (Hug, 2009), (Michelet et al., 2010), (Camara et al., 2010), (Dupuy-Chessa et al., 2011), (Priego-Roche, 2011), (Pernin et al., 2012), (Mandran et al., 2013), (Cornax, 2014), (Benkaouar, 2015). THEDRE a été soumis dans une conférence internationale. La communication porte sur le processus global et sa capacité à tracer une activité de recherche avec des indicateurs.. Depuis 7 ans, cette méthode est enseignée dans le cadre de deux écoles doctorales à Grenoble. Elle a été présentée dans des ateliers lors de conférences et dans d'autres laboratoires (LIP6 2015, LIUM 2016). THEDRE a été évaluée lors de 2 expérimentations, l'une sur le langage de description du processus, évalué par des experts en génie logiciel, l'autre sur les différents guides de conception, de conduite et de suivi d'expérimentation avec des doctorants et des chercheurs en RICH. Ces expérimentations ont validé la formalisation du processus, le méta-modèle associé, un dictionnaire des concepts et une notation graphique.

4.5.2 Maturité de THEDRE suivant Promote

D'après la manière dont THEDRE a été construite et évaluée, on peut estimer que le processus a *une procédure de validation et des résultats*. Cette description est semi-formelle : elle est spécifiée comme un langage spécifique à un domaine avec un dictionnaire des concepts, une syntaxe abstraite et une syntaxe concrète. Dans l'évaluation de la maturité d'une méthode, Promote propose aussi d'évaluer sa diffusion, autrement dit son adoption dans des situations concrètes suffisamment nombreuses. Cette approche est très spécifique de l'univers du développement informatique, dans lequel des millions d'ingénieurs utilisent des méthodes et en discutent sur des forums et où une large diffusion peut compter des dizaines de millions de pages Web. Elle semble moins pertinente pour ce qui est d'une méthode de conduite de recherche, en raison de la taille de la communauté et des habitudes de partage d'information en son sein.

Promote évalue aussi l'existence de métriques de bonne application du SDPM. Les différents indicateurs proposés dans THEDRE visant à assurer la qualité de la recherche et sa traçabilité, ces indicateurs permettent aussi de vérifier la bonne application de la méthode. Au sens de Promote, cette caractéristique correspond à une *procédure formelle de validation*.

4.6. Flexibilité dans THEDRE

4.6.1 Description de la flexibilité dans THEDRE

THEDRE propose d'avoir des blocs dépendants mais les tâches à l'intérieur des blocs ne sont pas systématiquement dépendantes. Certaines des tâches peuvent être menées en parallèle. L'organisation des tâches est libre, l'utilisateur du modèle de processus THEDRE choisit l'ordonnancement entre les tâches. Ensuite, il est présenté avec plusieurs niveaux de détails (sous-processus, blocs, tâches) ce qui apporte au processus un degré de flexibilité.

4.6.2 Caractérisation de la flexibilité dans THEDRE avec Promote

L'organisation de certaines tâches est laissée libre, mais elles sont obligatoires sans alternatives. THEDRE doit donc être réalisé complètement et n'offre pas de variants. De plus, il ne propose aucune procédure permettant son extension (ou sa réduction), par exemple pour ajouter de nouvelles activités. THEDRE n'est pas polymorphique : le processus est présenté comme une succession d'activités et ne peut pas être représenté en le centrant sur les artefacts produits. Ce polymorphisme pourrait être atteint en outillant le processus, mais le bénéfice reste à évaluer.

Nous l'avons vu, THEDRE offre des guides, un dictionnaire des concepts et des modèles de documents. L'ensemble de ces éléments assure au moins en partie des possibilités *d'apprentissage*, mais surtout fournit des éléments de *réutilisation*.

L'autre élément de flexibilité de THEDRE est sa granularabilité : le processus de THEDRE est décrit à plusieurs niveaux (sous-processus, blocs et tâches). On peut dire que THEDRE est au *troisième niveau de la granularabilité* : il offre un document principal avec une description à plus de deux niveaux de granularabilité.

5. Vers une fertilisation croisée entre processus de conduite de la recherche et ingénierie des processus

Nous venons d'appliquer Promote au domaine des processus de conduite de recherche en RICH. C'est une domaine très différent de son champ d'application initial. Nous souhaitons étudier si Promote peut être enrichie grâce à cette application originale. Par ailleurs, du point de vue de la conduite de la recherche, il était aussi intéressant de s'interroger sur l'intérêt de l'utilisation d'une approche d'ingénierie des processus.

5.1. Enrichissement de Promote

Promote sort de cet exercice avec un nouveau potentiel d'évolution pour l'évaluation des processus.

Au niveau de la notion d'artefact : dans Promote, un artefact est tout élément produit ou transformé par l'homme, ce qui était censé inclure la production de connaissances sur les SDPM. Promote évaluait la distensibilité des modèles de processus justement pour vérifier si les connaissances acquises pouvaient être capitalisées, par exemple comme nouvelles stratégies, activités ou éléments réutilisables. Or pour THEDRE, le substantif « artefact » n'est pas utilisé car sa polysémie pose des problèmes de compréhension entre le domaine de l'informatique et de l'analyse des données présente dans un des sous-processus de THEDRE. L'axe des artefacts pourra être étendu pour préciser leur nature (guide, indicateurs, …). De même que le sous-axe des livrables pourra être précisé avec les concepts proposés dans THEDRE tels que les protocoles ou les données expérimentales.

Au niveau du guidage et de l'apprentissage, Promote n'est pas supposé prendre en compte que ce qui est dit par les auteurs d'un SDPM. C'est ce qui conduit par

exemple à dire que les méthodes Agiles offrent peu de guidage, puisque celui-ci réside dans le champ cognitif des experts (e.g. les Scrum Masters) et non dans le SDPM. La réflexion apportée par THEDRE éclaire différents besoins : l'adjonction de *stratégies d'apprentissage* (dictionnaire des termes, méthodologie d'évaluation des SDPM,…), de *procédures d'extension* pour le compléter ou l'amender, et peut-être à terme la possibilité d'enrichir l'évaluation des SDPM avec ce qui se réalise dans ses instanciations.

Par ailleurs, nous avons vu que la nécessité d'un paradigme épistémologique pose la question de la validité et de la valeur de la connaissance. Nous pouvons porter ces concepts dans le domaine des SDPM: quelle sont la valeur et la validité de la solution réalisée, en prenant ces termes au sens du Worth-Centered Design de Cockton (Cockton, 2004). Dans sa version actuelle, Promote ne comporte pas de dimension s'intéressant au produit en tant que tel. Un tel axe pourrait questionner comment un modèle de processus de développement aide l'équipe qui l'emploie à favoriser la qualité – logicielle ou ergonomique – l'innovation et la créativité, l'adéquation aux besoins, l'étude de l'existant ou encore l'utilisabilité.

Enfin nous avons identifié une limite sur la mesure de la maturité d'un méthode : l'évaluation sera la même quel que soit le temps qu'il lui a fallu pour conquérir son public. Nous pourrions envisager d'ajouter ce critère à Promote.

Evolutions suggérées de Promote
Etendre l'axe des artefacts pour préciser la nature des artefacts et des livrables (données, guides …)
Intégrer des stratégies d'apprentissage dans le cadre de pratique métier
Intégrer des procédures d'extension
Intégrer des méthodes d'évaluation
Ajouter des axes relatifs à la valeur et à la validité du produit et du processus
Ajouter l'aspect temporel à l'évaluation de la maturité de processus

Tableau 2 : Synthèse des évolutions de Promote

5.2. Enrichissement de THEDRE

La caractérisation de THEDRE a montré que 24 des 36 axes de la taxonomie ont pu être appliquées directement. Promote est pleinement applicable à ce contexte.

Parmi ces 24 axes, neuf axes ont mis en évidence des améliorations pour THEDRE, par exemple en précisant mieux les rôles externes, en explicitant le niveau d'expertise attendu pour l'équipe qui mettrait en œuvre cette méthode, mais surtout en apportant une plus grande flexibilité dans le modèle de processus.

A l'inverse, six axes se sont avérés difficiles à utiliser en l'état.

Nous avons ainsi constaté que l'approche ascendante ou descendante devrait, pour traiter d'un processus de conduite de recherche, être centré sur la question du paradigme épistémologique. La question du centrage sur l'usage n'est pas apparue pertinente dans ce contexte, de même celle du type d'application, qui aurait gagné à

être formulé en termes de domaine d'utilisation. Les questions des risques et de maturité des besoins ne correspondaient à rien dans le contexte de THEDRE. Enfin, si les quantités de produits sont importantes dans une approche de développement itératif et incrémental, la question semblait sensiblement moins critique pour une conduite de recherche, où rien n'est dit sur ces aspects.

Evolutions de THEDRE
Améliorer la définition des rôles des acteurs dans le processus
Expliciter le niveau d'expertise attendue
Améliorer la flexibilité du processus

Tableau 3 : Synthèse des évolutions suggérées de THEDRE

5.3. Vers une taxonomie des processus de conduite de la recherche

Etant donnée la pertinence de Promote pour l'analyse de THEDRE, Promote pourrait facilement être adaptée pour fournir une taxonomie des processus de conduite de la recherche. Certains sous-axes seraient ainsi supprimés ou modifiés. D'autres pourraient être ajoutés. Ainsi en RICH, mesurer la maturité attendue des besoins ou évaluer si l'approche est centrée sur l'usage pour un processus de conduite de recherche n'a pas d'intérêt. Ces sous-axes peuvent être supprimés pour un processus de RICH. Nous pouvons aussi raffiner certains axes : la mesure de la diffusion telle qu'elle est présentée dans Promote (diffusion Internet, nombre de livres ou d'articles) n'est pas tout à fait satisfaisante : le public potentiel pour un méthode de RICH n'est pas aussi important que pour une méthode de développement logiciel. Il est donc difficile d'utiliser les mêmes ordres de grandeur lors d'une évaluation. De plus, le sous-axe traite de la diffusion de la parole sur une méthode et non de la diffusion de la méthode elle-même, et elle ne prend pas en compte l'ancienneté de la méthode. Ce point mérite d'être aussi ajouté en sous-axe pour la RICH. Le type d'application de Promote ne semble pas difficile à faire évoluer en cadre d'utilisation préconisé pour pouvoir prendre en compte les domaines de recherche concernés par un processus de RICH. Enfin, plutôt que d'étudier le type d'approche (descendante, ascendante ou composite), la manière d'aborder un projet de recherche pourra s'exprimer à travers un sous-axe relatif au paradigme épistémologique.

Un axe à ajouter pour la RICH est relatif à la validité et la valeur de la contribution scientifique et des outils activables. L'étude de THEDRE avec Promote a montré un manque dans ce domaine qu'il serait important de combler. Il est en particulier nécessaire d'étudier la qualité des données produites lors des expérimentations afin de garantir la validité de la recherche.

Modifications à apporter	**Axes ou sous-axes**
Supprimer	Maturité, collaboration / collaboration externe / centrée usage, Usage recommandé / risques et maturité des besoins, Artefacts / quantité(s)
Modifier	Cycle / approche avec comme approche le paradigme

	épistémologique, Usage recommandé / type d'application en domaines de recherche concernés, Maturité / diffusion (changer la mesure)
Ajouter	Maturité / Diffusion / ancienneté, axes Validité et Valeur

Tableau 4 : Synthèse des modifications à apporter à Promote

6. Conclusion et perspectives

Cet article présente l'évaluation d'un processus de conduite de la RICH à l'aide d'une taxonomie des processus des SDPM, Promote. L'évaluation de THEDRE a permis de mettre en évidence certains points faibles du processus. Mais cette analyse a aussi permis de mettre en évidence les apports et les limites de Promote pour caractériser des processus de conduite de la recherche, mais aussi de de SDPM. Ainsi des extensions de Promote ont été identifiées. Enfin ce travail permet de faire émerger une taxonomie spécifique aux processus de conduite de la RICH. Il doit être approfondi afin de finaliser cette taxonomie. Les axes ou sous-axes à ajouter ou à modifier doivent être étudiés afin de proposer des graduations pertinentes. Il en va de même pour les extensions de Promote. Enfin THEDRE, suivant une démarche d'amélioration continue, est en constante évolution. L'une de ces évolutions devra considérer le manque de flexibilité qui a été mis en avant par Promote.

Bibliographie

Avenier, M.-J., and Thomas, C. (2015). Finding one's way around various methodological guidelines for doing rigorous case studies: A comparison of four epistemological frameworks. Systèmes Inf. Manag. *20*, 61–98.

Benkaouar, W. (2015). Des Robots Compagnons avec du Style: Vers de la Plasticité en Interaction Social Humain-Robot.

Camara, F., Demumieux, R., Calvary, G., and Mandran, N. (2010). Cocoon, un système de recommandation sensible au contexte : analyse de la valeur par une étude qualitative. In Actes de La Conférence Ergo 'IA 2010, (ACM), pp. 211–218.

Céret, E. (2014). Flexibilité des processus de développement à la conception et à l'exécution : application à la plasticité des Interfaces Homme-Machine. University of Grenoble.

Céret, E., Dupuy-Chessa, S., Calvary, G., Front, A., and Rieu, D. (2013a). A taxonomy of design methods process models. Inf. Softw. Technol. Elsevier *55*, 795–821.

Céret, E., Calvary, G., and Dupuy-Chessa, S. (2013b). Flexibility in MDE for scaling up from simple applications to real case studies: illustration on a Nuclear Power Plant. In 25ème Conférence Francophone Sur l'Interaction Homme-Machine, IHM'13, (Bordeaux, France: ACM), pp. 33–42.

Cockton, G. (2004). From Quality in Use to Value in the World. In CHI '04 Extended Abstracts on Human Factors in Computing Systems, (New York, NY, USA: ACM), pp. 1287–1290.

Constantine, L.L., Biddle, R., and Noble, J. (2003). Usage-Centered Design and Software Engineering: Models for Integration. In ICSE Workshop on SE-HCI'03, pp. 106–113.

Cook, J.E., and Wolf, A.L. (1999). Software process validation: quantitatively measuring the correspondence of a process to a model. ACM Trans. Softw. Eng. Methodol. TOSEM *8*,

147–176.
Cornax, M.C. (2014). Amélioration Continue de Chorégraphie de Services: Conception et Diagnostic basés sur les Modèles.Thèse Université Grenoble Alpes.
De Vries, E.J. (2007). Rigorously Relevant Action Research in Information Systems. In ECIS, pp. 1493–1504.
Deming, W.E. (1952). Elementary Principles of the Statistical Control of Quality (Nippon Kagaku Gijutsu Remmei, Tokyo, Japon).
Drechsler, A., and Hevner, A. (2016). A four-cycle model of IS design science research: capturing the dynamic nature of IS artifact design. In Breakthroughs and Emerging Insights from Ongoing Design Science Projects: Research-in-Progress Papers and Poster Presentations from the 11th International Conference on Design Science Research in Information Systems and Technology (DESRIST) 2016. Canada, May, (DESRIST 2016)
Dupuy-Chessa, S., Mandran, N., Godet-Bar, G., and Rieu, D. (2011). A case study for improving a collaborative design process. In Engineering Methods in the Service-Oriented Context, (Springer), pp. 97–101.
Gregor, S., and Hevner, A.R. (2013). Positioning and presenting design science research for maximum impact. MIS Q. *37*, 337–355.
Harrison, N.B., and Coplien, J.O. (1996). Patterns of Productive Software Organizations. Bell Labs Tech. J. 11 138–145.
Hevner, A.R. (2007). A three cycle view of design science research. Scand. J. Inf. Syst. *19*, 4.
Howell, D.C., Rogier, M., Yzerbyt, V., and Bestgen, Y. (2007). Statistical Methods in Human Sciences. Boeck.
Hug, C. (2009). Méthode, modèles et outil pour la méta-modélisation des processus d'ingénierie de systèmes d'information. Thèse, Université de Grenoble-Alpes.
Jambon, F. (2009). User evaluation of mobile devices: in-situ versus laboratory experiments. Int. J. Mob. Hum. Comput. Interact. IJMHCI *1*, 56–71.
Jean-Daubias, S. (2004). De l'intégration de chercheurs, d'experts, d'enseignants et d'apprenants à la conception d'EIAH. In Technologies de l'Information et de La Connaissance Dans l'Enseignement Supérieur et de l'Industrie, pp. 290–297.
Mandran, N. (2017). THEDRE : méthode de conduite de la recherche. "Traceable Human Experiment Design Research." Thèse, Université de Grenoble-Alpes.
Mandran, N., Dupuy-Chessa, S., Front, A., and Rieu, D. (2013). Démarche centrée utilisateur pour une ingénierie des langages de modélisation de qualité. Ingénierie Systèmes Inf. *18*, 65–93.
Martin, R.C. (2003). Agile software development: principles, patterns, and practices (Prentice Hall PTR).
Michelet, S., Luengo, V., Adam, J.-M., and Mandran, N. (2010). How to take into account different problem solving modalities for doing a diagnosis? Experiment and results. In ITS 2010: 10th International Conference on Intelligent Tutoring Systems : Bridges to Learning, V. Aleven, J. Kay, and J. Mostow, eds. (Pittsburg, USA: Springer-Heidelberg), pp. 380–383.
Paille, P., and Mucchielli, A. (2011). L'analyse qualitative en sciences humaine et sociales (Armand Colin).
Peffers, K., Tuunanen, T., Gengler, C.E., Rossi, M., Hui, W., Virtanen, V., and Bragge, J. (2006). The design science research process: a model for producing and presenting information systems research. In Proceedings of the First International Conference on Design Science Research in Information Systems and Technology (DESRIST 2006), pp. 83–106.
Pernin, J.-P., Michau, F., Mandran, N., and Mariais, C. (2012). ScenLRPG, a Board Game for the Collaborative Design of GBL Scenarios: Qualitative Analysis of an Experiment. In Proceedings of the 6th European Conference on Games Based Learning, (Academic Publishing Limited), pp. 384–392.

Priego-Roche, L.-M. (2011). Modélisation intentionnelle et organisationnelle des systèmes d'information dans les organisations virtuelles. Thèse Université Grenoble Alpes.

Rolland, C. (2005). L'ingénierie des méthodes : une visite guidée A Guided Tour of Method Engineering.

Sein, M., Henfridsson, O., Purao, S., Rossi, M., and Lindgren, R. (2011). Action design research.

Simon, H.A. (2004). Les Sciences de l'artificiel (Paris: Folio).

Wang, F., and Hannafin, M.J. (2005). Design-based research and technology-enhanced learning environments. Educ. Technol. Res. Dev. *53*, 5–23.

Amélioration des méthodes de conduite de projets Big Data : retour d'expérience de pilotes industriels multi-sectoriels

Christophe Ponsard [1], **Mounir Touzani** [2], **Annick Majchrowski** [1]

1. *CETIC - Centre de recherche, Gosselies, Belgique*
   {christophe.ponsard,annick.majchrowski}@cetic.be
2. *Académie de Toulouse, Toulouse, France*
   mounir.touzani@ac-toulouse.fr

RÉSUMÉ. *Afin de mener à bien leurs activités, les entreprises sont de plus en plus confrontées au défi de traiter des quantités croissantes de données provenant de dépôts numériques, d'applications d'entreprise, de réseaux de capteurs... Bien qu'un large éventail de solutions techniques soit disponible pour traiter ces données massives (Big Data), beaucoup d'entreprises peinent à les déployer en raison d'un manque de maturité lié à leur gestion. Cet article propose une guidance en la matière. Il s'ancre dans des méthodes documentées dans la littérature, trouvant leurs racines dans les projets de fouille de données. Nous avons également mené une série de pilotes Big Data dans différents domaines (IT, médical, sciences de la vie, spatial) qui nous ont permis de dégager un retour d'expérience et un guide pratique pour la conduite d'un projet Big Data. Ceci permet d'exploiter au mieux les méthodologies disponibles afin de traiter les problématiques relatives à la collecte des exigences, l'exploration et la préparation des nouvelles données, le phasage itératif de l'implantation de la solution et une montée en maturité.*

ABSTRACT. *Nowadays, in order to successfully run their business, companies are facing the challenge to process ever increasing amounts of data generated from digital repositories, enterprise applications, sensors networks... Although a wide range of technical solutions are available to deal with such Big Data, many companies fail to deploy them actually because a lack of maturity in process and management challenges. This paper aims at providing guidance in those matter. We report about lessons learnt when deploying a series of Big Data pilots in different domains. We provide feedback and some practical guidelines on how to organise and manage a project based on available methodologies, covering topics like requirements gathering, data understanding, iterative project execution and raising the level of maturity.*

MOTS-CLÉS : *Gestion de projet, processus d'adoption, méthodes agiles, modélisation des données, Big Data, données massives, étude de cas*

KEYWORDS: *Projet Management, Adoption Process, Agile Methods, Big Data, Case Study*

1. Introduction

Notre monde est actuellement en train de vivre une explosion de l'information. De nombreuses statistiques attestent de la montée en puissance du phénomène Big Data. Par exemple, il est souvent rapporté que 90% des données dans le monde ont été produites seulement ces deux dernières années et que le volume des données créé par les entreprises double tous les 1,2 années (Rot, 2015).

Les organisations perçoivent bien le grand potentiel que les technologies Big Data peuvent leur apporter pour améliorer leur performance, et dans le cas des entreprises, pour accroître leur avantage compétitif. La facilité de collecter et stocker les données, combinée avec la disponibilité d'outils technologiques de stockage et d'analyse à grande échelle (notamment les bases de données NoSQL, MapReduce, Hadoop) a incité un certain nombre d'entre elles à démarrer des projets Big Data.

Les caractéristiques et défis posés par le Big Data sont souvent présentés au moyen d'une série de mots en "V" au Volume déjà mentionné, s'ajoutent notamment la Variété (diversité de formats structurés ou non), la Vélocité (aspect temps-réel du traitement), la Véracité (qualité des données), la Visualisation (afin de les interpréter facilement) et la Valeur (pour en tirer un revenu) (Mauro *et al.*, 2016).

Cependant, le constat est que la plupart des organisations ne parviennent toujours pas obtenir le dernier "V", c'est-à-dire produire une réelle valeur ajoutée à partir de leurs données. Un rapport de 2013 portant sur 300 entreprises Big Data a révélé que 55% des projets Big Data se sont terminés prématurément et que beaucoup n'ont que partiellement atteint leurs objectifs (Kelly, Kaskade, 2013). Ceci est confirmé par une étude en ligne conduite par Gartner en juillet 2016, qui a rapporté que de nombreuses entreprises restent bloquées au stade du projet pilote et que seulement 15% des projets Big Data ont été effectivement déployés en production (Gartner, 2016).

En examinant la cause de tels échecs, il apparaît que le facteur principal n'est en réalité pas lié à la dimension technique, mais plutôt aux processus et aux aspects humains qui s'avèrent être aussi importants (Gao *et al.*, 2015). Un examen de la littérature révèle qu'actuellement, de nombreux articles se concentrent encore énormément sur la dimension technique, en particulier l'utilisation d'algorithmes qui permettent de réaliser des analyses approfondies, et que beaucoup moins d'attention est portée aux méthodes et aux outils qui pourraient aider les équipes à mener efficacement des projets Big Data à terme (J. Saltz, Shamshurin, 2016).

Il existe toutefois quelques travaux récents dans ce domaine, notamment en matière d'identification des facteurs clés de succès des projets Big Data (J. S. Saltz, 2015), aussi bien sur des problèmes de gestion de projet (Corea, 2016) que sur la manière dont les équipes s'organisent pour réaliser des projets Big Data, en pointant cependant l'absence de standard en la matière (J. Saltz, Shamshurin, 2016).

Notre article se situe dans la lignée de ces travaux et a pour objectif d'apporter des recommandations concrètes aux entreprises engagées dans un processus d'adoption de solution Big Data. A travers ce travail, nous souhaitons apporter quelques éléments

de réponses à des questions telles que :

– Comment pouvons-nous être sûrs que le Big Data pourrait nous aider ?
– Quelles personnes devraient être impliquées et à quel moment ?
– Quelles sont les étapes clefs auxquelles il faut être attentif ?
– Est-ce que mon projet est sur la bonne trajectoire pour aboutir ?

Notre contribution se veut de nature pratique et s'appuie sur un ensemble de projets pilotes couvrant différents domaines (sciences de la vie, santé, espace, infrastructures informatiques). Ces pilotes sont répartis sur deux ans et sont réalisés dans le cadre d'un projet global commun, réalisé en Belgique. Le processus suivi est similaire et renforcé progressivement. Les travaux rapportés sont basés sur les quatre premiers pilotes et quatre autres sont en phase de planification.

Ce document est structuré comme suit. La section 2 donne une typologie des principales catégories de projets Big Data. La section 3 passe ensuite en revue les principales méthodologies concernant le déploiement du Big Data. Dans la section 4, nous présentons la méthodologie suivie pour mener nos projets pilotes et dégager une guidance méthodologique. Nous mettons l'accent sur les facteurs clés de succès du déploiement d'une solution Big Data. La section 5 détaille notre retour d'expérience en donnant des recommandations ciblant des étapes particulièrement importantes. Enfin, la section 6 tire quelques conclusions et extensions que nous envisageons de mener dans la suite de nos projets pilotes.

2. Typologie des méthodes d'analyse de données massives

L'analyse de données ("Data Analytics") est un concept multidisciplinaire qui peut être défini comme les moyens permettant d'acquérir des données depuis de sources diverses, de les traiter afin de découvrir des relations qui les relient et mettre des résultats à disposition des parties prenantes (H. Chen *et al.*, 2012). L'application de ces techniques par des entreprises ("Business Analytics") leur permet de mieux comprendre leur niveau de performance et de procéder à des améliorations. Trois catégories complémentaires d'analyse peuvent être distinguées et combinées pour atteindre les objectifs de compréhension des données et d'aide à la décision.

– *L'analyse descriptive* permet d'investiguer le passé afin de répondre à la question "Que s'est-il passé ?". Elle repose sur un ensemble de techniques permettant d'examiner les données pour comprendre et analyser les performances de l'entreprise. Il s'agit notamment de l'analyse statistique ainsi que de méthodes de classification et de catégorisation. Elle comprend également le diagnostic pour répondre à la question : "Pourquoi est-ce arrivé ?", afin de comprendre les raisons des événements qui se sont produits dans le passé.

– *L'analyse prédictive* est tournée vers l'avenir et essaie de répondre aux questions "Que va-t-il se passer ?" et "Pourquoi cela risque-t-il de se produire ?". Elle utilise un ensemble de techniques d'analyse des données actuelles et passées pour découvrir ce qui est le plus susceptible de se produire (ou non). Les approches utilisées ici sont

principalement basées sur l'exploration de données et l'apprentissage automatique ("machine learning")

– *L'analyse prescriptive* examine également l'avenir, mais permet de mettre l'accent sur les recommandations et conseils afin de répondre aux questions "Que dois-je faire ?" et "Pourquoi devrais-je le faire ?". Les techniques spécifiques qui sont utiles ici, relèvent de l'optimisation, de la simulation, des systèmes de règles métier voire de systèmes experts permettant de proposer des actions contre les risques connus ou identifiés via l'analyse prédictive.

3. Revue des méthodes et processus existants

Cette section passe en revue les méthodes et processus existants pour la mise en œuvre de projets Big Data. Elle souligne certaines forces et limitations connues. Nous commençons par présenter des méthodes héritées du domaine de la fouille de données (Data Mining ou DM) et de l'informatique décisionnelle (Business Intelligence ou BI) avant d'envisager des approches plus spécifiques au Big Data avec une attention particulière aux méthodes agiles. Enfin certaines méthodes complémentaires inspirées d'approches plus cognitives ou de gestion de la maturité seront également envisagées.

3.1. Méthodes liées à la fouille de données et l'informatique décisionnelle

La fouille de données a été développée dans le courant des années '90 avec pour objectif d'extraire des données à partir d'informations structurées (bases de données) pour découvrir des facteurs clés de l'entreprise à une échelle relativement petite. Le Big Data, quant à lui, opère aussi sur des données non structurées, sur une plus grande échelle et vise à dégager des indicateurs à vocations prédictives. Cependant, un point commun aux deux types d'approches est qu'en termes de processus, il est nécessaire de mettre en place une coopération étroite entre les experts techniques (données) et les experts métiers (Hoppen, 2015). De nombreuses méthodologies et modèles de processus ont été développés pour la fouille de données et la découverte de connaissances (Mariscal *et al.*, 2010).

L'informatique décisionnelle s'est également développée dans les années '90 et vise essentiellement à produire des indicateurs clé de performance (en anglais KPI : Key Performance Indicator) sous forme de tableaux de bord. Les techniques s'appuient sur des données structurées et ne nécessitent que peu d'intelligence dans les traitements. Le Big Data permet d'élargir le champ de la BI aux données moins structurées. Inversement, la BI apparaît comme un prérequis permettant de mesurer précisément ce qu'on désire améliorer tandis que les techniques Big Data apportent des possibilités d'analyse prédictive (Halper, 2014).

L'approche séminale en matière de fouille de données est KDD (Knowledge Discovery in Database). Elle a été raffinée en plusieurs autres approches (SEMMA, Two Crows, etc.) avant d'être standardisée par CRISP-DM (Cross Industry Standard Process for Data Mining, ou processus standard pour la fouille de données, en français)

(Shearer, 2000). Cette méthode est décrite dans la figure 1. Elle est composée de six phases, chacune étant décomposée en sous-étapes. Le processus n'est pas linéaire, mais plutôt organisé comme un cycle global avec généralement des revues entre les phases. CRISP-DM a été largement utilisé depuis 20 ans, non seulement pour la fouille de données, mais aussi pour l'analyse prédictive et des projets Big Data.

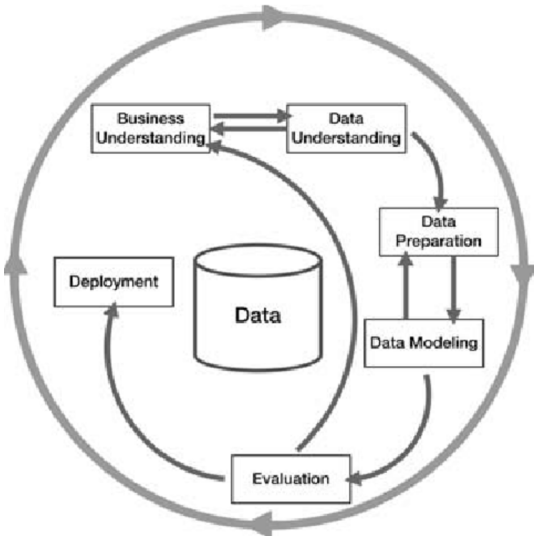


Figure 1. La méthode CRISP-DM

CRISP-DM et les méthodes similaires souffrent toutefois des problèmes suivants :

– elles ne fournissent pas une bonne vision du management du point de vue communication ainsi qu'au niveau de la connaissance et sur les aspects des projets.

– elles manquent d'une certaine forme de maturité au niveau du modèle pour permettre de mettre en évidence des étapes et des jalons plus importants, qui peuvent être améliorés progressivement.

– en dépit de la normalisation, elles ne sont pas très largement connues des entreprises qui peuvent donc difficilement les adopter pour mieux gérer la valeur de leurs données.

3.2. Vers plus d'agilité

Les méthodes agiles sont des méthodes itératives qui répondent au manifeste agile dont les principes mettent l'interaction avec le client, l'adaptation aux changements et la production de valeur au centre du processus de développement (Alliance, 2001). Initialement développées pour le développement de logiciels, ces principes peuvent également répondre plus largement et en particulier à l'analyse des données afin de fournir une meilleure guidance en particulier pour aboutir à la production de valeur. Une évolution de KDD et CRISP-DM vers l'agilité est assurée par la méthode AgileKDD (Nascimento, Oliveira, 2012). Celle-ci est basée sur le cycle de vie OpenUP

qui répond aux principes du Manifeste Agile (Balduino, 2007). Les projets sont divisés en "sprints" planifiés avec des délais fixes, habituellement de quelques semaines. Dans chaque sprint, les équipes doivent produire de la valeur ajoutée aux parties prenantes de manière prédictive et démontrable.

Bien que les méthodes agiles semblent en adéquation avec les besoins, le déploiement de telles méthodes pour le Big Data peut se heurter à une résistance, tout comme c'est le cas dans le domaine du développement logiciel. C'est en particulier le cas dans les organisations de plus grande taille qui sont habituées à des processus assez rigides plus aisés à planifier. Une enquête a révélé que tout comme pour le logiciel, les entreprises ont tendance à accepter des méthodes agiles pour les projets Big Data de plus petite envergure, moins complexes et ayant peu d'exigences liées à la sécurité. Il s'agit aussi généralement d'organisations plus flexibles. En dehors de ces cas, l'approche préférée reste l'approche planifiée (Franková *et al.*, 2016).

3.3. Méthodes spécifiques pour le Big Data

La méthode AABA (Architecture-centric Agile Big data Analytics) répond aux défis techniques et organisationnels de Big Data (H.-M. Chen *et al.*, 2016). La méthode intègre à la fois une méthode de conception du système Big Data (BDD) et une architecture AAA (Architecture-centric Agile Analytics). Elle est centrée sur le modèle DevOps et orientée vers la découverte efficace et livraison continue de valeur.

La méthode a été validée sur 11 études de cas dans différents domaines notamment en marketing, télécom et santé. Sur cette base, elle a émis les recommandations suivantes :

1. les analystes et experts en données doivent être impliqués tôt dans le processus
2. un soutien continu aux activités d'architecture est nécessaire
3. des pics d'efforts en mode agile permettent de faire face aux évolutions rapides des technologies et des exigences
4. la définition d'une architecture de référence permet une plus grande flexibilité.
5. les boucles de rétroaction permettent de traiter les exigences non fonctionnelles telles que la performance, la disponibilité et la sécurité, mais aussi pour disposer d'un retour rapide des clients à propos d'exigences émergentes.

Parallèlement, *Stampede* est une méthode proposée par IBM à ses clients. Son principal objectif est d'encourager les entreprises et les aider à démarrer plus rapidement, afin de générer de la valeur à partir du Big Data. La méthode s'appuie sur la mise à disposition de ressources d'experts à un coût permettant d'aider les entreprises à se lancer dans le Big Data, dans le cadre d'un projet pilote bien défini (IBM, 2013). La méthode, illustrée à la figure 2, s'appuie notamment sur un atelier d'une demi-journée permettant de définir le projet Big Data, d'identifier l'infrastructure nécessaire, d'établir un plan de travail mais surtout et avant tout d'établir la valeur pour l'entreprise. L'exécution du pilote est généralement répartie sur 12 semaines et réalisée de manière agile avec un jalon important vers la 9ème semaine.

Figure 2. Méthode Stampede d'IBM

Des tentatives ont également été menées pour développer un *modèle de maturité de capacité de type CMM* (Capability Maturity Model) pour les processus de gestion des données scientifiques, dans le but de soutenir l'évaluation et l'amélioration de ces processus (Crowston, 2010) (Nott, 2014). Un tel modèle décrit les principaux types de processus ainsi que les pratiques nécessaires à une gestion efficace. Un CMM caractérise les organisations au moyen d'un niveau de maturité qui représente leur capacité à exécuter des processus de façon fiable. Une échelle classique sur 5 niveaux est typiquement utilisée à la fois dans (Crowston, 2010) et (Nott, 2014). Le premier utilise les niveaux standard allant de "défini" à "optimisé" tandis que le second utilise une nomenclature plus spécifique allant de "ad hoc" à "breakaway". Le tableau 1 en détaille les principaux critères qui concernent la place de la donnée dans la stratégie métier, le type d'analyse de données utilisée, l'alignement de l'infrastructure IT, ainsi que des aspects de culture et de gouvernance.

Tableau 1. Modèle de maturité de Nott et Betteridge (IBM)

Niveau	Ad hoc	Fondateur	Compétitif	Différentiateur	Libérateur (Breakaway)
Stratégie métier	Utilisation de reporting standard. Big Data juste évoqué	Identification d'un ROI lié aux données	Exploitation des données encouragée	Réalisation d'un avantage compétitif	Innovation métier conduite par les données
Analyse de données	Limité au passé	Détection d'événement	Prédiction de certaines probabilité d'évolution	Optimisation des décisions	Optimisation et automation possible de certains processus
Alignement IT	Pas d'architecture cohérente ni unifiée	Framework architectural présent mais adapté au Big Data	Définition de patrons architecturaux pour le Big Data	Architecture définie et standardisée pour la plupart des "V"	Architecture totalement alignée avec les besoins Big Data
Culture et gouvernance	Largement basé sur des individualités	Gestion fragmentaire, résistance au changement	Définition de politiques et de procédures, adoption partielle	Adoption large, utilisation quotidienne	Adoption et mise en oeuvre généralisée

3.4. Approches complémentaires

Sensemaking est également une approche itérative, mais en rapport avec les processus cognitifs réalisés par les humains afin de se construire une représentation mentale de l'information pour atteindre l'objectif visé. Elle met l'accent sur les défis de la modélisation et de l'analyse en intégrant les modèles cognitifs afin d'analyser les

caractéristiques des données et de détailler les activités des utilisateurs (Lau *et al.*, 2014).

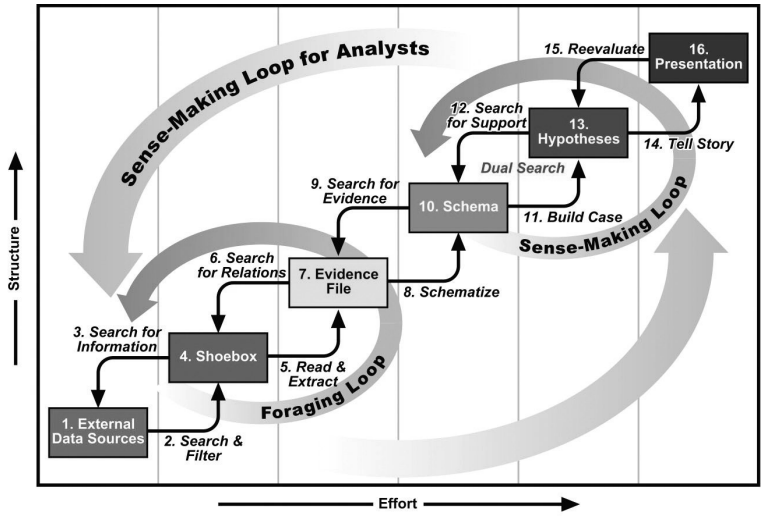


Figure 3. Méthode SenseMaking

De nombreux facteurs clés de succès, guides pratiques et listes de contrôle des risques ont été également publiés, principalement dans les blogs pour les directeurs des systèmes d'information, p. ex. (Bedos, 2015). Une classification systématique des facteurs critiques de succès a été proposée par (Gao *et al.*, 2015) en utilisant trois dimensions clés : les personnes, les processus et la technologie. Celle-ci a été étendue ensuite par (J. Saltz, Shamshurin, 2016) pour traiter aussi des dimensions de l'outillage et de la gouvernance. Les principaux facteurs clés sont les suivants :

– pour les données : la qualité, la sécurité, le niveau de structure des données

– pour la gouvernance : une direction, une organisation bien définie, une culture axée sur les données

– pour les objectifs : la valeur de l'entreprise identifiée (KPI), la rentabilité, une taille de projet réaliste

– pour les processus : l'agilité, la conduite de changement, la maturité, la volumétrie des données

– pour l'équipe : des compétences en ingénierie des données, la multidisciplinarité

– pour les outils : des infrastructures informatiques, le stockage, la capacité de visualisation des données, le suivi des performances

4. Processus global suivi pour développer et valider la méthode

4.1. Aperçu du processus global

L'objectif global de notre projet est d'élaborer une méthode systématique pour aider les entreprises à valider les avantages potentiels d'une solution Big Data. Le processus global est représenté dans la figure 4, il est guidé par huit pilotes successifs qui sont utilisés pour affiner la méthode et rendre plus technique les briques disponibles à travers l'infrastructure proposée. Le résultat final attendu est de fournir un service reproductible de manière fiable aux entreprises ayant de tels besoins.

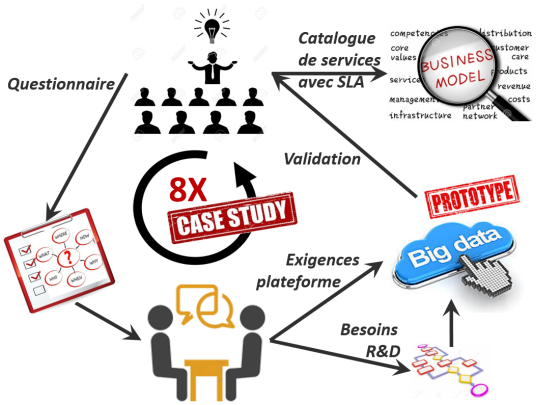


Figure 4. Développement itératif de la méthode et de l'infrastructure

La méthode choisie est fortement inspirée des méthodes et processus décrits dans la section 3 :

– le point de départ a été Stampede grâce à la plate-forme d'IBM. Les principaux aspects retenus à partir des méthodes sont : l'atelier initial avec toutes les parties prenantes, la focalisation réaliste et un moteur de la valeur d'entreprise constant.

– Pour faire face à un manque de matériel de référence, nous avons défini un modèle de processus basé sur CRISP-DM qui est largement documenté.

– les pilotes sont exécutés de manière agile, étant donné les disponibilités des experts (chercheurs universitaires), et planifiés sur des périodes plus longues que dans Stampede : 3-6 mois au lieu de 12-16 semaines. L'approche populaire SCRUM a été également utilisée car elle met l'accent sur la collaboration, le fonctionnement du logiciel, l'autogestion de l'équipe et la flexibilité pour s'adapter aux réalités de l'entreprise (Scrum Alliance, 2016).

4.2. Caractérisation des projets pilotes

Les différents pilotes sont gardés confidentiels. Le tableau 2 en donne néanmoins les principales caractéristiques exprimées notamment au moyen des 3 premier "V" du Big Data ainsi que de la typologie

Tableau 2. Principales caractéristiques des 4 premiers projets pilotes

#	Domaine	Volume	Vélocité	Variété	Caractérisation
1	Sciences de la vie	20 Go/analyse, 2 To/semaine	Haute (à paralléliser)	Données métier et de traçabilité (ex. agro-alimentaire)	Essentiellement descriptive, au niveau de la qualité des produits
2	Spatial	Maintenance infrastructure sol Galileo	Moyenne	Haute: messages, logs	Maintenance prédictive de matériel coûteux. Fiabilité 99.8%
3	Santé	900 lits sur 3 sites	Temps-réel	Nombreuses sources, formats divers	Analyse prédictive et prescriptive pour réduire la morbidité. Confidentialité.
4	Maintenance IT	Environ 3000 serveurs	Haute (événements, logs,...)	Temps-réel	Analyse prédictive pour maîtriser les coûts de maintenance

4.3. Schéma général appliqué au sein de chaque pilote

La méthodologie qui a émergé sur base des méthodologies existantes et sur base des itérations sur nos 4 pilotes se compose de trois phases suivantes :

Phase 1. Contexte et sensibilisation au Big Data.

Dans cette phase d'introduction, une ou plusieurs réunions sont organisées avec l'organisation participante. Une introduction générale est donnée sur les concepts du Big Data. La plate-forme mise à disposition est présentée, de même que quelques applications représentatives dans différents domaines (éventuellement avec déjà un focus sur le domaine de l'organisation). Les principaux défis et les étapes clés de la mise en œuvre sont également exposés. Lors des interactions, le niveau de maturité du client et certains facteurs de risque peuvent déjà être vérifiés (par exemple, l'implication de la direction, le niveau d'expertise interne, la formulation d'objectifs assez clairs).

Phase 2. Compréhension de l'entreprise et du cas d'utilisation.

Cette phase est largement alignée avec la première phase de CRISP-DM présentée à la section 3.1. Son objectif est d'identifier les besoins et problèmes pour lesquels une solution de type Big Data est envisagée. Il est aussi important de formuler un ou plusieurs cas d'utilisation qui peuvent démontrer l'apport de valeur à partir des données collectées et traitées. Il s'agit d'une phase très importante et des outils méthodologiques concrets pour aider à la conduire sont détaillés dans la section 4.

Phase 3. Mise en œuvre d'un pilote pour un service ou un produit.

Dans cette phase, les activités suivantes sont menées de manière agile :

– *Compréhension des données* : analyser les données pour en extraire les sous-ensembles les plus intéressants et assurer une bonne qualité des données.

– *Préparation des données* : sélectionner les données pertinentes, les nettoyer, les étendre et les formater selon les besoins.

– *Modélisation* : sélectionner une technique de modélisation spécifique (par exemple, arbre de décision ou réseaux de neurones). Le modèle est alors construit puis testé au niveau de sa précision et sa généralité (mais pas encore en relation avec les besoins de l'entreprise). Le respect des hypothèses de modélisation est également vérifié. A partir des résultats, les paramètres du modèle peuvent être revus ou d'autres techniques complémentaires peuvent être utilisées.

– *Évaluation* : évaluer dans quelle mesure le modèle répond aux objectifs de l'entreprise, en utilisant des données réalistes ou même réelles.

– *Déploiement* : transférer la solution validée à l'environnement de production et veiller à ce que l'utilisateur puisse l'utiliser (par exemple, au moyen de bons outils de visualisation et d'un tableau de bord). Les activités de surveillance de performance et de précision sont également mises en place.

5. Retour d'expérience et recommandations

Dans cette section, nous présentons nos principaux retours d'expérience ainsi que des recommandations méthodologiques permettant d'augmenter les chances de succès d'un projet de déploiement Big Data.

5.1. Définition d'objectifs progressifs et dont la valeur est mesurable

Par le déploiement d'une solution Big Data, une entreprise s'attend à gagner de la valeur de ses données. La façon de mesurer cette valeur doit être définie dès la phase de compréhension des données de l'entreprise, généralement en s'appuyant sur les indicateurs clés de performance (KPI). Ces KPI doivent déjà être clairement définis par l'entreprise et celle-ci doit être déjà en mesure de les mesurer.

Sur cette base, différentes stratégies d'amélioration peuvent être identifiées et discutées pour aboutir à la sélection d'un bon projet pilote de mise en œuvre. Dans ce processus de sélection, l'écart avec la situation actuelle et le niveau de maturité doivent également être pris en considération. Il est plus sûr de garder un premier projet avec des objectifs assez modestes que de risquer l'échec en visant un projet trop complexe, même s'il pourrait apporter plus de valeur. Une fois que ce premier projet pilote réussi, d'autres étapes peuvent être planifiées pour mettre en place des traitements plus complexes amenant plus de valeur à l'entreprise.

5.2. Du réactif au préventif puis au prédictif

Dans plusieurs domaines, il est intéressant de mettre en place un schéma permettant d'évoluer vers une réaction immédiate à des caractéristiques identifiées à travers les données, vers plus d'intelligence afin d'anticiper des situations indésirables, voire

les prévenir suffisamment pour pouvoir les éviter. Nous donnons ici deux illustrations respectivement dans le domaine de la maintenance et de la santé.

Étude de cas du domaine de la maintenance informatique. En matière de maintenance, un KPI est le coût total d'appartenance (TCO - Total Cost of Ownership). Celui-ci inclut le coût d'achat, de maintenance et de réparation en cas de panne. Différentes stratégies peuvent être envisagées :

– *réagir* simplement aux problèmes après la survenue d'une panne. Ceci se traduit par un coût généralement important car il faut réagir rapidement afin de minimiser le temps d'indisponibilité. Par ailleurs toute indisponibilité a un impact négatif en termes d'image voire de pénalité si un SLA (Service Level Agreement) a été violé.

– *anticiper* leur occurrence sur la base de l'observation du système. Des stratégies simples peuvent être mises en place, par exemple déclencher des alertes quand un stockage approche d'un seuil proche de la capacité maximale. Ceci ne permet cependant pas de prévoir des pannes résultants d'enchaînements complexes d'événements.

– *tenter de prédire* les problèmes sur base d'historique connu et d'observation du système. C'est à ce niveau que des techniques d'analyse de données permettent de mettre en évidence des relations de cause à effet entre des parties du système qui, en cascade, peuvent causer une indisponibilité. Par exemple l'application d'un correctif mal validé peut affecter un service qui peut lui-même paralyser un processus métier.

– *optimiser* l'étape ultime. Il faut veiller constamment à ce que le système opère dans des conditions optimales en éliminant les causes des pannes possibles à la source.

La solution prédictive est la meilleure à notre sens, mais elle ne devrait être envisagée que si l'étape préventive est réalisée. De même, les patrons temporels les plus fréquents doivent être identifiés et traités en premier, par exemple, les stockages risquent plus une saturation les jours où des sauvegardes sont effectuées, généralement de manière prévisible (fin de semaine ou fin de mois). Une anticipation permettrait d'éviter des interventions coûteuses, notamment le week-end.

Étude de cas dans le domaine des itinéraires de soins. En matière de soins de santé, les hôpitaux déploient de plus en plus des trajets de soins, définis comme une vision pluridisciplinaire du processus de traitement requis par un groupe de patients présentant la même pathologie avec un suivi clinique prévisible (Campbell *et al.*, 1998). Il peut par exemple s'agir d'un pontage cardiaque ou d'une chimiothérapie. Ceci permet de non seulement de réduire la variabilité des processus cliniques mais aussi d'améliorer la qualité et mieux en maîtriser les coûts (Dam, 2013). La mise en place d'itinéraires permet aussi de faire une analyse plus riche des données produites : on peut ainsi détecter des patients ayant un profil qui pourrait impacter la qualité de leur traitement (par exemple lié à une autre pathologie dont il souffre ou des intolérances).

L'automatisation de ces analyses est d'autant plus importante que le suivi est généralement pluridisciplinaire et que certaines interactions peuvent être complexes à appréhender par un seul spécialiste et peuvent donc potentiellement échapper à l'analyse humaine. Ceci est particulièrement critique dans la cas de traitement tel que le

cas des chimiothérapies où le respect du temps et des doses est important. Un indicateur médical défini en la matière est le RDI (Relative Dose Indicator). Dans le cas du cancer du sein, il a été montré que la dégradation de cet indicateur avait un impact direct sur la courbe de rémission (Piccart *et al.*, 2000). La surveillance du RDI par le système et l'analyse prédictive des facteurs qui l'impacte est donc primordiale.

5.3. Guidance dans la phase de compréhension du métier et des données

Cette phase est critique pour le succès du projet car l'objectif n'est pas seulement d'aboutir à une compréhension des besoins et des données disponibles mais aussi de mettre en place le noyau de personnes qui sera porteuse de la suite du projet. A cette fin, on recommande de l'organiser sur la base d'un ou plusieurs ateliers impliquant un responsable commercial, l'analyste des données et l'architecte SI. D'autres experts peuvent aussi être impliqués plus ponctuellement, par exemple, le responsable de la sécurité informatique peut être consulté pour valider à un stade précoce les problèmes possibles de sécurité ou confidentialité. A la fois le système actuel et l'évolution future du système d'information doivent être considérés. Afin de mener cette phase, une liste de contrôles utiles est représentée à la figure 5.

Figure 5. Compréhension de l'entreprise et des cas d'utilisation

Pour soutenir l'organisation d'une manière efficace, des outils spécifiques de ces ateliers sont décrits à la section 5. A la fin de cette étape, une planification de projet est également définie.

La tenue d'un atelier exige de prêter attention à de nombreuses questions tout en concentrant la discussion sur les plus pertinentes. A cet égard, un questionnaire peut fournir un soutien efficace à la fois comme préparation avant l'atelier et comme une

liste de contrôle (check-list) pendant celui-ci. Le tableau 3 illustre quelques questions utiles à la compréhension des données à traiter.

Tableau 3. Quelques questions d'atelier sur les données

Q.UD.1	Quelles sont les sources de données et les types de données utilisés dans vos processus métier actuels ?
Q.UD.2	Quels outils/applicatifs sont utilisés pour traiter vos processus métier actuels ?
Q.UD.3	Vos processus métier actuels effectuent-ils un traitement complexe des données ?
Q.UD.4	Quelle est la disponibilité de vos données ? Que se passe-t-il si les données ne sont pas disponibles ?
Q.UD.5	Des utilisateurs autres ont-ils un droit d'accès différent sur vos données ?
Q.UD.6	Vos données contiennent-elles des informations sensibles (par exemple, des données personnelles ou confidentielles de l'entreprise) ?
Q.UD.7	Quelles sont les conséquences de l'altération des données ?
Q.UD.8	Connaissez-vous le niveau de qualité de vos données ?

5.4. Utilisation de notations pour la modélisation

L'utilisation de notation de modélisation est utile comme outil pour inventorier les données, comprendre leur structure et comprendre les différents flux d'information. Il ne faut cependant pas la confondre avec l'étape technique de modélisation qui est ultérieure. Pendant les ateliers, un tableau blanc peut être utilisé pour esquisser des modèles dans un mode collaboratif avec les participants.

Selon notre expérience, les modèles de flux de données aident à comprendre quel processus génère, modifie, stocke ou extrait des données. Les modèles d'entités-relations (ou diagrammes de classe ou ontologies) aident à capturer la structure du domaine

Par contre, les cas d'utilisation doivent être évités car ils se focalisent sur une fonctionnalité spécifique mais ne permettent pas de mettre en évidence les liens entre les données. Ils ne peuvent donc pas fournir une image globale du problème.

5.5. Mise en place de points de contrôle

L'approche agile permet au processus d'être flexible et incrémental sur les activités. Avant de commencer une activité, il faut cependant disposer d'un minimum de résultats des étapes précédentes. Dans ce but, le tableau 4 reprend quelques contrôles à consulter au démarrage d'une activité.

6. Conclusion et perspectives

Dans cet article, nous avons décrit comment aborder les défis et les risques liés au déploiement d'une solution Big Data au sein d'organisations et en particulier d'entreprises souhaitant s'appuyer sur cette technologie pour soutenir leur développement. Sur base de différentes méthodes et études déjà rapportées dans la littérature, nous

Tableau 4. Liste (partielle) de vérification de la préparation à l'évaluation

R.EV.1	Êtes-vous capable de comprendre/utiliser les résultats des modèles ?
R.EV.2	Est-ce que les résultats du modèle vous semblent pertinents d'un point de vue purement logique ?
R.EV.3	Y a-t-il des incohérences apparentes qui méritent d'être approfondies ?
R.EV.4	D'après votre première vision, les résultats semblent-ils répondre au métier de votre organisation ?

avons élaboré de manière itérative une méthode adaptée à nos besoins en y intégrant des retours d'expérience de plusieurs pilotes. Au delà de cette méthode qui continue à évoluer au fil de projets pilotes, notre principale contribution est centrée sur le processus suivi pour mettre en place un projet Big Data qui maximise les chances de succès et qui s'adapte aux besoins de l'organisation cible. Nous proposons en outre une série de recommandations soutenant cette mise en œuvre. Bien que centrée sur quelques pilotes, notre approche se veut donc générale et permet aux personnes confrontées aux même défis de disposer de briques méthodologiques utiles pour déployer efficacement un projet Big Data et bien en gérer les difficultés et pièges.

Jusqu'à présent, nous nous sommes focalisés davantage sur les phases de découverte et de compréhension des données. Dans la suite de nos travaux, nous explorerons plus en détails la phase d'exécution du projet au fur et à mesure que nos projets pilotes auront atteint leur terme ou des jalons importants.

Remerciements

Ce travail a été financé en partie par le projet PIT Big Data de la Région wallonne (no 7481). Nous remercions nos partenaires d'avoir partagé leur cas d'étude.

Bibliographie

Alliance A. (2001). *Agile Manifesto.* http://agilemanifesto.org.

Balduino R. (2007). *Overview of OpenUP.* https://www.eclipse.org/epf/general/OpenUP.pdf.

Bedos T. (2015). *5 key things to make big data analytics work in any business.* http://www.cio.com.au/article/591129/5-key-things-make-big-data-analytics-work-any-business.

Campbell H., Hotchkiss R., Bradshaw N., Porteous M. (1998). Integrated care pathways. *British Medical Journal*, p. 133-137.

Chen H., Chiang R. H. L., Storey V. C. (2012, décembre). Business intelligence and analytics: From big data to big impact. *MIS Q.*, vol. 36, n° 4.

Chen H.-M., Kazman R., Haziyev S. (2016). Agile big data analytics development: An architecture-centric approach. In *Proc. hicss'16, hawaii, usa.*

Corea F. (2016). *Big data analytics: A management perspective* (1st éd.). Springer Publishing.

Crowston K. (2010). A capability maturity model for scientific data management.

Dam P. A. van. (2013). A dynamic clinical pathway for the treatment of patients with early breast cancer is a tool for better cancer care: implementation and prospective analysis between 2002–2010. *World Journal of Surgical Oncology*, vol. 11, nº 1, p. 70.

Franková P., Drahošová M., Balco P. (2016). Agile project management approach and its use in big data management. *Procedia Computer Science*, vol. 83.

Gao J., Koronios A., Selle S. (2015). Towards A Process View on Critical Success Factors in Big Data Analytics Projects. In *Amcis.*

Gartner. (2016). *Investment in big data is up but fewer organizations plan to invest.* http://www.gartner.com.

Halper F. (2014). *Predictive Analytics for Business Advantage.* The Data Warehousing Institute Best Practices Report, TDWI.

Hoppen J. (2015). *7 characteristics to differentiate BI, Data Mining and Big Data.* https://aquare.la/articles/2015/05/01/7-characteristics-differentiate-bi-data-mining-big-data.

IBM. (2013). *Stampede.* http://www.ibmbigdatahub.com/tag/1252.

Kelly J., Kaskade J. (2013). *CIOs & Big Data: What Your IT Team Wants You to Know.* http://blog.infochimps.com/2013/01/24/cios-big-data.

Lau L., Yang-Turner F., Karacapilidis N. (2014). Requirements for big data analytics supporting decision making: A sensemaking perspective. In *Mastering data-intensive collaboration and decision making.* Springer Science & Business Media.

Mariscal G., Marban O., Fernandez C. (2010). A survey of data mining and knowledge discovery process models and methodologies. *Knowledge Eng. Review*, vol. 25, nº 2, p. 137-166.

Mauro A. D., Greco M., Grimaldi M. (2016, 04 04). A formal definition of big data based on its essential features. *Library Review*, vol. 65, nº 3, p. 122-135.

Nascimento G. S. do, Oliveira A. A. de. (2012). An agile knowledge discovery in databases software process. In *Data and knowledge engineering: Third international conference, icdke, wuyishan, fujian, china, nov. 21-23.* Springer Berlin Heidelberg.

Nott C. (2014). *Big Data & Analytics Maturity Model.* http://www.ibmbigdatahub.com/blog/big-data-analytics-maturity-model.

Piccart M., Biganzoli L., Di Leo A. (2000, Apr). The impact of chemotherapy dose density and dose intensity on breast cancer outcome: what have we learned? *Eur J Cancer.*, vol. 36.

Rot E. (2015). *How Much Data Will You Have in 3 Years?* http://www.sisense.com/blog/much-data-will-3-years.

Saltz J., Shamshurin I. (2016). Big Data Team Process Methodologies: A Literature Review and the Identification of Key Factors for a Project's Success. In *Proc. IEEE International Conference on Big Data.*

Saltz J. S. (2015). The need for new processes, methodologies and tools to support big data teams and improve big data project effectiveness. In *IEEE int. conf. on big data.*

Scrum Alliance. (2016). *What is scrum? an agile framework for completing complex projects.* https://www.scrumalliance.org/why-scrum.

Shearer C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing*, vol. 5, nº 4.

Utilisation de la Méthode DEA pour l'Évaluation des Performances des Processus Métier

Mourad Bouneffa, Benoît Becquet, Henri Basson

École des Ingénieurs du Littoral Côte d'Opale
Laboratoire d'Informatique Signal et Image de la Côte d'Opale
50, rue Ferdinand Buisson - BP 719, 62228 CALAIS CEDEX*, FRANCE*
mourad.bouneffa,benoit.becquet,henri.basson@eilco-ulco.fr

RÉSUMÉ. *Les organisations mettent en œuvre de nombreux processus métiers qui peuvent être formalisés et explicitement déployés à l'aide de technologies telles que SOA (Service Oriented Architecture); simplement modélisés à l'aide de nombreux formalismes telles que BPMN (OMG, 2011) ou EPC (Fellmann* et al.*, 2013) ou existants à travers des notes de procédures standardisées. Nous nous intéressons à la mesure de la performance des processus métier indépendamment de la manière dont ils sont mis en oeuvre. Nous proposons une approche consistant à recenser les différents éléments d'un processus métier et à le traduire en une sorte de boite noire dont les entrées et les sorties serviront de données de base à un processus d'évaluation de la performance. Ce processus utilise la méthode d'enveloppement de données ou DEA. Nous appliquons notre approche aux processus de gestion des stages dans différentes composantes d'une université.*

ABSTRACT. *To achieve their business objectives, organisations implement various Business Processes. These may be precisely defined, formalized and deployed using some specific technologies such as SOA (Service Oriented Architectures). In somer organisations, processes are just defined using specific formalisms such as BPMN (OMG, 2011) or EPC (Fellmann* et al.*, 2013) without an explicit relationship or link with IT elements implementing them. Finally in some other organisations processes are not formalized nor defined at all but exist by means of standardized documents and notes. In this paper, we deal with measuring the performance of Business Processes that may be or not formalized. Our approach is based on listing all the artefacts of a business process and then representing such a process as a black box in which inputs and outputs are relevant data for performance measurement using the Data Envelopment Analysis method. The approach is currently adopted to measure the performance of internships management as implemented by various departments of a french university.*

MOTS-CLÉS : *Processus Métier, Evaluation des Performances, méthode DEA*

KEYWORDS: *Business Processes, Efficiency evaluation, DEA method*

1. Introduction

La notion de processus métier est fondamentale pour la maîtrise du développement, du déploiement, de l'évaluation et de l'évolution des différents systèmes assurant la bonne gestion d'une organisation. En effet, ces processus représentent une sorte de cartographie fonctionnelle et organisationnelle de l'organisation (Glasson *et al.*, 1994). Leur formalisation facilite la mise en oeuvre de nombreuses activités du cycle de vie d'un système organisationnel telle que la mise en oeuvre d'un nouveau système informatique de gestion intégré, l'adoption d'une nouvelle technologie informatique telle que l'utilisation de dispositifs mobiles ou l'adoption de nouvelles règlementations ou normes à l'image des nouvelles règles de comptabilité (Haworth, Pietron, 2006), etc. Dans certaines organisations, le système de gestion est entièrement conçu, développé et déployé autour de la formalisation des processus métier. Dans ces organisations, la formalisation des processus métier en utilisant des langages standards tels que BPMN (OMG, 2011) est suivie par une projection de ces processus en termes de macro programmes représentant l'orchestration de services web qui constituent les briques logicielles de base constituant le système informatique support des activités de l'organisation (Recker, Mendling, 2006 ; Doux *et al.*, 2009 ; Mazanek, Hanus, 2011 ; Ouyang *et al.*, 2009). Cette approche qui consiste à partir de la modélisation des processus pour produire un système informatique vu comme une orchestration de services web est appelée Service Oriented Architecture (Newcomer, Lomow, 2005) et vise à maîtriser le cycle de vie d'un système organisationnel de la formalisation des processus jusqu'à leur déploiement. En général, cette approche est basée sur un modèle de procédé itératif et incrémental dans le sens où le cycle de vie est une suite d'itérations incluant la modélisation des processus, leur implémentation, leur déploiement, leur évaluation et finalement leur évolution (Bouneffa *et al.*, 2016). Malheureusement, tous les processus métiers ne sont pas forcément définis ni déployés de cette manière dans toutes les organisations. Dans certaines organisations, il existe bien des processus métiers formalisés mais cela est dicté par des besoins de conformité à des normes et certifications de la qualité telles que les différentes normes ISO. Dans ce cas, il est rare que les acteurs de l'organisation se servent des processus ainsi formalisés comme un moyen d'assurer le cycle de vie des applications qui sont mises en oeuvre pour l'informatisation des différentes procédures de l'organisation. La dernière catégorie d'organisations sont celles qui ne disposent d'aucune formalisation de leurs processus. Ces derniers existent par le biais des procédures connues à travers des pratiques courantes et supportées par des documents plus ou moins normalisés.

Dans ce papier, nous nous intéressons à la problématique de l'évaluation de la performance des processus métiers. Nous ne posons aucun préalable sur la manière dont les processus doivent être modélisés, implémentés ou déployés. Le cadre d'évaluation inclut toute sorte d'organisations quelle que soit la manière dont les processus sont mis en œuvre.

Notre approche d'évaluation est essentiellement basée sur l'utilisation de la méthode DEA ou Data Envelopment Analysis (Badillo, 1999 ; Diagne, 2006 ; Seiford, 1997 ; Wei, 2001) qui est une méthode de mesure de la performance multi-facteurs

déjà utilisée dans de nombreux domaines telles que l'évaluation des performances des agences bancaires (LA VILLARMOIS, 1999) ou la mesure de la performance des centres d'approvisionnements (logistique amont) (Cavaignac, Villesèque-Dubus, 2014), etc. Le processus que nous mettons en œuvre et qui est décrit dans ce papier consiste principalement à traduire un processus métier en une formulation d'un problème d'évaluation de performances selon la méthode DEA. En d'autres termes, coder le processus en termes d'inputs ou ressources utilisées par le processus et d'outputs ou résultats retournés par le processus. La méthode DEA consistera donc à réaliser une évaluation de plusieurs processus effectuant des tâches similaires pour déterminer les meilleurs en terme de performance et calculer alors pour chaque processus les écarts par rapport aux meilleurs. Le décideur peut alors optimiser le processus en allouant moins de ressources tout en gardant les mêmes résultats ou en gardant les mêmes ressources en augmentant les résultats. La traduction d'un processus en le formulant selon un problème d'optimisation selon la méthode DEA passe par une étape intermédiaire qui est la production d'une représentation synthétique d'un processus. Cette représentation synthétique va consister à recenser les différentes entrées, sorties, ressources d'un processus. Selon la nature du processus cette étape peut se faire de façon semi-automatique ou manuellement.

La suite du papier est structurée comme suit. La section 2 décrit les éléments principaux de la méthode DEA. Dans la section 3, nous décrivons les éléments principaux de notre approche. Nous y décrivons particulièrement le processus d'extraction des entrées sorties de la méthode DEA à partir d'éléments descriptifs d'un processus métier. La section 4 montre l'application de notre approche à des processus métier issus de la gestion des stages dans les différentes composantes d'une université. La section 5 conclut le papier en revenant sur ses apports et les perspectives qu'il ouvre.

2. Éléments de la méthode DEA

La méthode DEA ou Data Envelopment Analysis est une méthode d'évaluation des performances des unités organisationnelles appelées DMUs ou Decision Making Units. Cette méthode a été introduite par Charnes et al. (Charnes *et al.*, 1981) comme un moyen d'évaluer les performances ou plutôt l'efficience d'un programme fédéral américain d'allocation de ressources aux écoles. Le but étant de mesurer les écoles les plus efficientes en matière d'utilisation de ressources. L'efficience est mesurée par un ratio entre les ressources utilisées et les résultats obtenus. Par la suite, cette méthode a été étendue à de nombreux domaines incluant la gestion des hôpitaux publiques, les services sociaux, les agences d'un réseau bancaire, etc. Le principe de la méthode est assez simple. Elle considère tout système ou unité décisionnelle comme une sorte de boite noire consommant des ressources (appelées *inputs*) et produisant des prestations (appelées *outputs*). Une des utilisations de la méthode est de comparer plusieurs unités organisationnelles ou DMUs pour en déterminer les meilleures. Ces dernières formeront une sorte de frontière d'efficience. Toutes les unités en dessous de la frontière sont réputées avoir une marge de progression qu'il sera aisée de calculer. Le terme enveloppement de données vient de cette frontière qui permet d'envelopper tous les

cas possibles de combinaisons d'inputs et de outputs. Ayant déterminé les marges de progression, les décideurs peuvent alors jouer sur les inputs en les diminuant tout en gardant le même niveau de prestation (outputs) ou agir sur les outputs en les augmentant tout en gardant le même niveau d'inputs.

Pour illustrer la méthode DEA et son utilisation, nous allons adopter un exemple très simple. Prenons le cas d'un ensemble de départements d'une université disposant chacun de moyens propres pour réaliser les inscriptions des étudiants. Pour des raisons de simplification nous considérons un seul type d'inputs (le nombre d'équivalents temps plein d'employés) et un seul type d'outputs (le nombre d'étudiants inscrits par jour).

Tableau 1. Nombre d'inscriptions par département d'une université

	Input	Output
	Nombre de Secrétaires	Nombre d'Inscrits
Département A	2	1
Département B	3	4
Département C	5	5
Département D	4	7
Département E	6	7

Le tableau 1 représente le rendement journalier des processus d'inscriptions des 5 départements d'une université. A partir de ce tableau plutôt simpliste, on peut déterminer l'enveloppe ou la frontière d'efficience selon deux hypothèses : l'hypothèse selon laquelle les organisations évoluent dans une situation de rendements d'échelle constants (modèle constant returns to scale ou CRS) et l'hypothèse de selon laquelle les organisations évoluent dans une situation de rendements d'échelle variables (modèle variable returns ro scale VRS). Les résultats de l'analyse selon ces deux modèles sont données respectivement par les figures 1 et 2.

Dans la figure 1, on remarquera que le département réalisant le meilleur rendement (le département B) détermine à lui seul l'enveloppe ou la frontière. En effet, comme on part du postula que les rendements d'échelle sont constants, cela veut dire qu'il n'y a pas d'autres éléments intrinsèques aux unités qui influent sur le rendement. En d'autres termes, une secrétaire devrait avoir le même rendement quelque soit le département et cela ne dépend pas des spécificités des départements ou de certaines contraintes particulières. Donc l'enveloppe est une droite qui passe par l'origine du repère et traverse le point correspondant au département B. Tous les points situés en dessous de cette droite ont une marge de progression. Ainsi, pour le point A, si l'on raisonne selon les $inputs$, le rendement de A ou son non efficience est obtenu en faisant le rapport du segment SA'/SA. On utilisant une règle on obtient ainsi la valeur 0.37 ou 37%. Cela signifie que A pourrait réduire ses effectifs de 62.5% (100-37) tout en gardant le même rendement.

Dans la figure 2, on se limite à tracer une courbe qui relie les meilleurs points et qui détermine une frontière qui englobe tous les autres points. Cette courbe qui

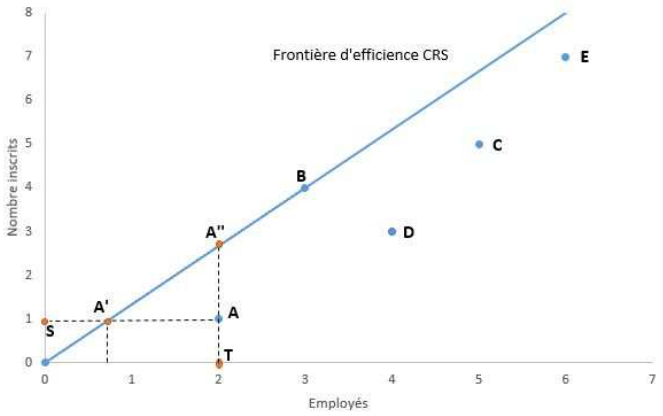


FIGURE *1. Frontière ou Enveloppe DEA selon le modèle CRS*

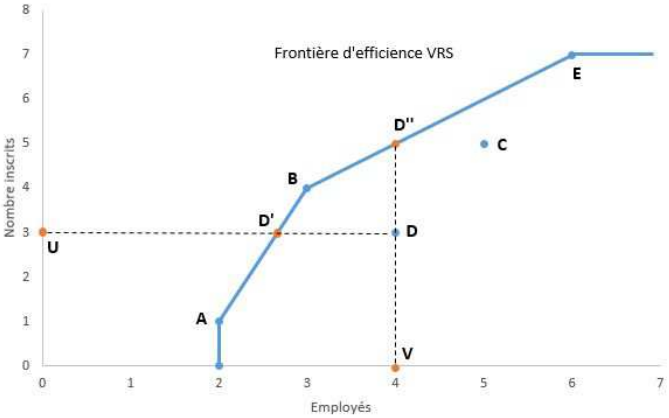


FIGURE *2. Frontière ou Enveloppe DEA selon le modèle VRS*

relie les points A, B et E représente donc l'enveloppe ou la frontière. Dans ce cas, pour le point D par exemple, et si on s'intéresse aux $outputs$, la non efficience est obtenue en faisant le rapport de la quantité D"D/DV. Cela donne une valeur de 60%. En d'autres termes D pourrait augmenter le nombre des inscriptions de 40% (100-60) tout en gardant le même nombre de personnel. Si par contre on considère les inputs, le rapport UD'/UD qui donne une valeur de 66,7% indique que D pourrait réduire le nombre de ses employés de 33,3% en gardant la même production.

Dans l'exemple, que nous venons d'expliciter, la méthode DEA considère un seul type d'$input$ et un seul type de $output$. En réalité, la méthode est généralisable à plusieurs $inputs$ et à plusieurs $outputs$. Dans ce cas, elle va consister à maximiser une quantité : $w_1 * output_1 + w_2 * output_2 + \ldots + w_n * output_{Nn}$ tout en minimisant la combinaison des inputs. Cela va revenir à un problème de programmation linéaire et il s'agira de calculer les poids $w_1, \ldots, w_n$. En effet, la méthode de DEA ne permet pas aux décideurs de fixer eux mêmes les valeurs des poids associés aux outputs et aux inputs mais les calcule. En effet, ces poids ont une valeur comprise entre 0 et 1. La meilleure unité obtient une valeur de 1 pour les poids et les autres unités obtiennent des valeurs inférieures. Ainsi, si une unité obtient 0.85 pour un facteur, cela s'interprète par le fait qu'elle a un manque à gagner de 25% en terme d'efficience. Supposant par, exemple, que pour les départements d'une université, on considère comme output le nombre des inscrits et le nombre de boursiers. Si un département obtient un poids de 0.85 pour les inscrits et 0.5 pour les boursiers cela s'interprète par le fait que ce département a un manque à gagner de 25% en terme de nombre d'étudiants à inscrire et de 50% pour le cas des étudiants boursiers.

3. Approche de l'évaluation des performances des processus métiers

La figure 3 résume les éléments principaux de notre approche. En entrée le processus métier est décrit par un certain nombre d'éléments que sont : les artefacts ou données d'entrées du processus, les ressources utilisées par le processus et les données ou artefacts de sorties du processus. Ces éléments peuvent être obtenus de plusieurs façons.

Dans le cas d'un processus formalisé à l'aide d'un langage tel que BPMN ou XPDL (Palmer, 2009) nous disposons, comme résultat de nos précédents travaux de moyens d'analyses lexico-syntaxique permettant d'extraire les éléments principaux d'un processus. Nous disposons également d'une ontologie (Bouneffa, Ahmad, 2013 ; HENDI *et al.*, 2016 ; Bouneffa *et al.*, 2016) permettant de représenter des informations d'ordre sémantique pouvant avoir un lien avec le domaine métier concerné par le processus. Une extraction des différents éléments caractérisant un processus est alors réalisée en appliquant des requêtes SPARQL (Prud'hommeaux, Seaborne, 2008) utilisant au niveau des critères des informations issues d'annotations sémantiques. Nous pouvons par exemple exprimer des requêtes du types : *extraire toutes les données d'entrées des activités effectuées par des acteurs qui sont des comptables*, etc.

Dans le cas ou le processus n'est pas formalisé, le seul moyen dont nous disposons est l'interview des acteurs clés du processus mais focalisée entièrement sur les entrées, les sorties et les ressources utilisées pour un processus donné sans rentrer dans les détails de spécification des activités du processus.

Le résultat est un modèle simplifié d'un processus métier qu'on appellera Représentation Synthétique du Processus ou RSP.

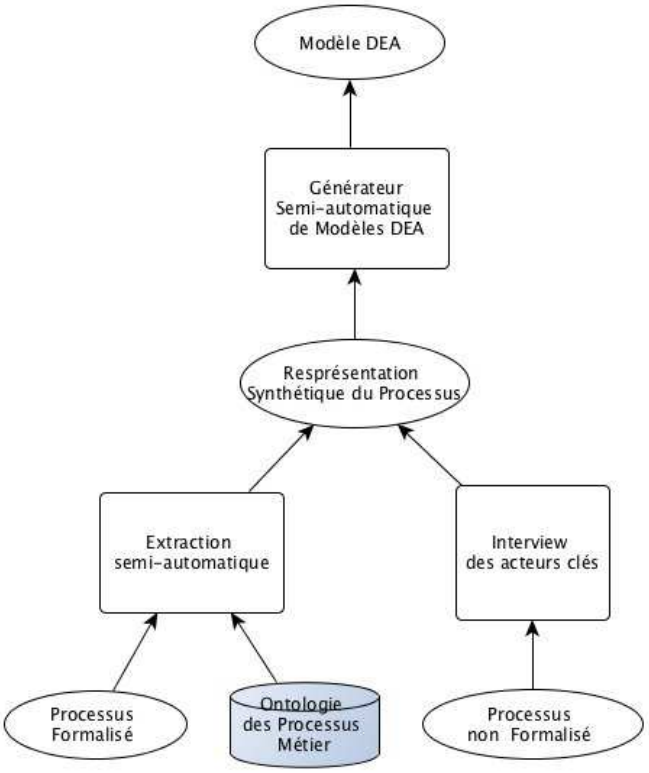


FIGURE *3. Extraction d'un modèle DEA à partir d'un processus métier*

La RSP est alors utilisée comme entrée d'un outil semi-automatique permettant d'extraire une modélisation du problème traité par le processus de départ sous une forme exploitable par une méthode d'évaluation des performances des entités économiques largement utilisée et ayant fait ses preuves dans le monde de l'économétrie et qui est la méthode DEA ou Data Envelopment Analysis. Notre but étant d'évaluer les performances de plusieurs unités exécutant un même processus avec éventuellement des variantes et des outils différents mais ayant les mêmes entrées et les mêmes sorties, cette méthode nous permet ainsi de définir une classification en terme de performances de ces unités et de déterminer ainsi les meilleurs pratiques, les écarts par rapport à ces meilleures pratiques de chaque unité implantant le processus.

Un outil de résolution de ce type de problèmes est alors utilisé pour l'évaluation Envelopment Analysis) ou Approche par Enveloppement de Données qui sera décrite dans la section suivante.

3.1. Structure d'une **RSP**

Une représentation synthétique d'un processus peut être vue comme un tuple : *t*=<*Entrées, Sorties, Map, Ressources, Caractéristiques*>.

– *Entrées :* l'ensemble des données d'entrée du processus. Si l'on prend le cas d'un processus de gestion commerciale, cela peut inclure les demandes de devis, les commandes clients, etc.

– *Sorties :* l'ensemble des données de sortie du processus. Pour un processus de gestion commerciale cela peut inclure les devis, les factures, les commandes, etc.

– *Map :* est un mapping entre les entrées et les sorties. C'est une relation de correspondance permettant de représenter un lien de causalité entre une entrée et une sortie. Dans le cas d'un processus commercial un map peut contenir les couples (demande de devis, devis), (commande client, facture), etc.

– *Ressources :* l'ensemble des ressources utilisées par le processus. Dans la majorité des cas, on se limitera aux ressources humaines mais cela peut inclure également les ressources informatiques, les moyens de communication, etc. S'il s'agit par exemple d'établir un benchmarking d'unités disposant toutes des mêmes ressources matérielles (informatiques par exemple), on peut ne pas prendre en compte cette ressource car elle est la même pour tout le monde.

– *Caractéristiques :* un ensemble de caractéristiques qui peuvent servir de support d'explication des résultats en matière d'analyse des performances. Les caractéristiques qui nous semblent intéressantes concernent la manière dont le processus est mis en oeuvre.

On distinguera notamment les caractéristiques suivantes :

– *Formalisé :* un processus est dit formalisé s'il a été complètement modélisé en utilisant un langage de modélisation des processus. Un tel langage peut-être BPMN, XPDL, ECP, etc.

– *Implémenté :* un processus est dit implémenté s'il existe une application informatique qui implémente ses différentes activités.

– *Monitoré :* un processus est dit monitoré s'il existe un moyen informatique permettant de récupérer des données sur son fonctionnement. Par exemple, dans un processus de gestion commerciale, on est capable d'avoir en temps réel le nombre de devis produits par heure ou par poste de travail, la durée moyenne de traitement d'une demande de devis, etc.

Un processus peut être juste *Implémenté* ou *Implémenté* et *Formalisé* ou bien *Implémenté*, *Formalisé* et *Monitoré*. Dans le cas où le processus est juste *Implémenté*, cela veut dire qu'il existe une ou plusieurs applications informatiques qui mettent en œuvre les différentes activités du processus. Dans le cas où il est *Implémenté* et *Formalisé* cela veut dire que non seulement il est existe une ou plusieurs applications informatiques qui déploient ses activités mais il existe également une formalisation à jour en terme d'un langage standard des différentes activités. Cependant, on ne formule aucune hypothèse sur le fait qu'il y ait un lien exploitable du point de vue infor-

matique entre la formalisation et l'implémentation. En d'autres termes, nous n'avons pas de mapping ou projection directe entre une activité et le composant informatique qui l'implémente. Un processus est dit *Monitoré* s'il existe un moyen informatique de superviser son fonctionnement. En d'autres termes, il est possible de récupérer des données sur les exécutions de chacune de ses activités comme le temps moyen d'exécution d'une activité, etc. Dans ce cas, nous supposons qu'il existe forcément un lien explicite entre les activités ou tâches du processus et les composants informatiques l'implémentant.

Les caractéristiques que nous venons de décrire vont nous servir comme support d'explication de certains résultats que la méthode DEA risque de fournir lors de l'évaluation des processus. Cependant, nous pensons pouvoir les convertir en termes de inputs de la méthode DEA. En effet, la production d'une formalisation d'un processus, le monitoring ou l'implémentation ont certainement un coût. Ce coût est donc forcément un facteur important dans la mesure de l'efficience économique. Cependant, dans l'état actuel de nos travaux nous n'avons pas encore intégré de procédures claires de calcul de coût pour chaque type de processus. Nous pensons que ce calcul nécessite un raffinement des caractéristiques et une étude empirique sur différentes organisations pour déterminer avec plus de précision les facteurs de ce type de coût.

3.2. Processus de construction d'un modèle DEA à partir de la RSP d'un processus

Nous avons donc proposé un procédé permettant de traduire un processus représenté par sa RSP en un modèle de problème DEA représenté par un ensemble d'inputs et de outputs. Ce procédé est schématisé par le pseudo algorithme du listing 1.

Listing 1 – TraductionProcessusEnDEA (entree : rsp une RSP sortie : dea un modele DEA)

```
1
2   Pour chaque sortie s de rsp.sorties
3                                   produire un output o de dea de type fonctionaggregation(s).
4                                    /* fonctionagregation peut etre nombre(), somme(), moyenne((), etc.
5                                    Elle est selectionnee de maniere interactive par l utilisateur */
6   Pour chaque ressource r de rsp.Ressources
7                               produire un input i de dea de type count(r).
8
9   Pour chaque map((e, s)) de rsp.Map reliant une entree e a une sortie s
10                                Proposer un output o de dea de type fonction(e,s)
11                                 /* fonction peut-etre un ratio par exemple nombre
12                                 de factures par nombre de demandes de
13                                 devis ou une fonction temporelle comme le temp moyen
14                                  separant les arrivees des
15                                 demandes de devis et la production des devis*/
```

4. Exemple d'Application

Afin d'évaluer notre approche, nous nous sommes intéressés au processus de ges-

tion des stages au sein de notre université. On comparera 3 entités n'ayant absolument pas le même formalisme. In fine, nous chercherons à établir une classification en termes de performances de ces unités. L'ensemble de cette démarche nous permettra de déterminer ainsi les meilleures pratiques et les écarts par rapport à ces meilleures pratiques. Les entités étudiées sont :

- Le service des stages de l'IUT.
- Le service stage du département EEA de l'université
- Le service stage de l'Ecole d'Ingénieurs.

Chaque service étudié gère les stages de manière différente, avec plus ou moins de formalismes. Pour parvenir à nos résultats nous avons donc mené des interviews afin d'obtenir les éléments nécessaires pour faire notre étude. Le premier constat que nous avons pu effectuer est qu'il était, pour le moment impossible de générer de façon automatique le modèle DEA et cela non pas par manque d'outillage mais surtout par le fait que dans la majorité des services les processus sont soit pas formalisés ou formalisés pour des besoins d'assurance qualité en utilisant des formalismes de types logigrammes adaptés à une lecture et compréhension par un gestionnaire. Pour le cas des logigrammes, nous avons chargé des étudiants de traduire certains en BPMN ce qui nous a permis d'expérimenter la génération des RSP et des modèles DEA. Cependant, nous n'avons effectué cette activité que pour très petite partie des logigrammes faisant partie du document d'assurance qualité.

Nous décrivons ci-dessous les processus de gestion de stage de chaque entité sélectionnée en précisant le niveau de formalisme pour chacun.

4.1. Description des processus de gestion de stage par entité de l'université

Au sein de l'école d'ingénieurs, les processus relatifs à la gestion des stages ont été formalisés par des procédures élaborées dans le cadre d'une certification qualité ISO 9001. Il n'existe pas de systèmes intégré dédié à la gestion des stages. En d'autres termes, les procédures sont utilisées comme un outil documentaire afin de comprendre le fonctionnement de la gestion des stages. Les outils utilisés pour l'implémentation sont un mélange de tableurs Excel, de procédures de gestion électronique de documents à travers l'extranet de l'école et de procédures manuelles utilisant le support papier. Si nous nous limitons à la procédure de validation d'un sujet de stage qui forme la phase amont au déroulement de stage, nous pouvons décrire le processus par le texte suivant :

Après avoir reçu des offres de stages qui émanent soit du service de gestion des stages, directement de l'entreprise ou par une recherche personnelle de l'étudiant, celui-ci doit faire valider sa demande de stage par le responsable pédagogique de l'année. Le responsable a pour mission de s'assurer que le stage répond bien aux critères fixés pour la nature du sujet et pour la durée. Une fois cette validation acquise, l'étudiant doit se rendre physiquement chez la secrétaire du service des stages afin d'y faire établir la convention de stage. Selon le pays d'accueil du stagiaire, la convention

adaptée sera établie et l'étudiant enverra sa convention pour validation à l'entreprise. Les conventions signées par les différentes parties sont ensuite remises aux parties intéressées, archivées à l'école et l'étudiant pourra réaliser son stage.

Pour l'école d'ingénieurs nous avons fait traduire par des étudiants la procédure décrite ci-dessus en un diagramme BPMN joint à l'annexe de ce papier. Une analyse lexico-syntaxique du diagramme BPMN nous a permis d'extraire les éléments nécessaire à l'élaboration de la RSP et du modèle DEA. Pour les cas du département EEA et de l'IUT le problème était encore plus compliqué puisqu'il n'y avait même pas de procédure formelle décrivant la gestion des stages. Le seul moyen d'extraire les éléments nécessaires était donc le recours aux interviews du personnel et des usagers (étudiants et quelques entreprises accueillant des stagiaires). Dans ces deux systèmes, il s'avère que la phase en amont du stage, à savoir la validation du sujet du stage, dépend du responsable pédagogique mais aucune démarche n'est explicitée.

4.2. Extraction du modèle DEA de l'exemple d'application

Comme nous l'avons décrit dans les sections précédentes, notre approche se concentre sur les entrées, sorties et ressources des systèmes. L'analyse des différentes entités de l'université nous a quand même permis de définir une RSP commune décrite par les éléments suivants ;

– *Entrées :* { offres de stages émanant des entreprises, demandes de validation de stages provenant des étudiants, }.

– *Sorties :* { conventions de stages signées pour la France, conventions de stages signées pour l'étranger, stages soutenues }.

– *Ressources :* { responsable pédagogique, secrétaire stage }.

– *Map :* { (demande de validation de stage, convention de stage), (convention de stage, stage soutenu), (offre de stage, convention de stage), (offre de stage, stage soutenu)) }.

Dans cette RSP commune nous n'avons pas représenté les caractéristiques car elles sont propres à chaque entité. Ainsi, au niveau de l'école d'ingénieurs le processus est formalisé alors que dans les autres entités il ne l'est pas. Nous ne pouvons pas dire que le processus est implémenté, ni monitoré dans chacune des structures car il n'existe ni application intégrée, ni orchestration de services web implémentant le processus.

Avant de dresser le tableau de chiffres représentant la modélisation de notre problème à l'aide de la méthode DEA , nous commençons par dresser un bilan chiffré des entrées, sorties et ressources (figure 4). Dans cette figure les ressources qui représentent les secrétaires et les enseignants chercheurs (responsables pédagogiques) sont exprimés en équivalent temps plein.

Nous avons décidé d'appliquer l'algorithme du listing 1 avec comme fonction d'agrégation $count$ ou $nombre$ appliquée sur les sorties et les ressources. Nous n'avons pas pris en compte les entrées mais il était possible de le faire pour obtenir des ratios

	Entrées			Sorties			Ressources	
	Offres Stages Entreprises	Offres Stages Etudiants	Demande validation de stage	Conventions France signées (CFR)	Conventions étranger signées (CETR)	Stages Soutenus (S_EFF)	Secrétaires (SCER)	Enseignants chercheurs (ENS_CH)
IUT	120	10	105	100	0	100	0,5	0,5
Département Informatique	70	10	55	50	0	50	0,25	0,25
Ecole d'Ingénieurs	250	80	310	170	130	300	0,5	0,5

FIGURE *4. Bilan annuel chiffré des entrées, sorties et ressources des entités de gestion des stages*

de type pourcentage de conventions de stages signées par offres de stages reçues, etc. La modélisation DEA va alors correspondre aux nombres de sorties qui formeront les *outputs* du modèle DEA et aux nombres de ressources représentant les *inputs* du modèle DEA.

En appliquant ces valeurs à un solveur DEA (le solveur win4deap [1]), nous avons d'abord calculé l'efficience des différentes unités. Comme expliqué précédemment l'organisation la plus efficience obtient un score de 1. Dans notre simulation c'est l'école d'ingénieurs qui obtient ce score. Les deux autres entités obtiennent chacune un score de 0,588 représentant donc un décalage de 1-0,588 soit 0,412 ou 41,2% par rapport à l'entité la plus efficiente. Deux approches sont alors possibles pour remédier au problème d'efficience de ces deux entités. Une approche par les inputs ayant comme résultat attendu la diminution des ressources allouées en gardant le même résultat (mêmes quantités de outputs) ou une approche par output consistant à augmenter les outputs en gardant les mêmes ressources. Une simulation orientée outputs et appliquée à l'IUT et le département EEA nous a permis de proposer une recommandation de passer le nombre de convention signées en France respectivement à 300 et à 150.

5. Conclusion

Nous avons présenté un processus permettant d'évaluer les performances des processus métier en utilisant une méthode issue de l'économétrie et qui est la méthode d'enveloppement de données ou DEA. Le but de ce papier est avant tout de proposer une approche qui puisse servir de base pour le développement de sortes de plugins qui seront associées aux outils d'aide à la mise en œuvre d'applications de modélisation, de développement, de déploiement et de monitoring de processus métiers dans les différentes organisations. Bien que nous ayons développer un outil permettant de produire une modélisation DEA des processus métiers formalisées à l'aide de langages tels que BPMN, notre approche ne suppose pas l'existence d'un tel outil et est

1. http://www.sigmdel.ca/aed-dea/install-en.html

applicable pour toute sorte de processus métier y compris ceux qui ne sont pas du tout formalisés. Nous proposons, en effet, un cadre d"évaluation qui se veut applicable à toutes les organisations incluant celles pour lesquels il n'y a pas encore eu un effort de formalisation des processus. Comme exemple de ce type d'organisation, nous avons choisi le cas de notre université et avons expérimenté notre approche pour évaluer la performance des services de gestion des stages étudiants de 3 de ses entités. Nous avons en effet, moyennant un petit effort de compréhension des processus mis en œuvre, pu effectuer une évaluation des performances de ces trois entités. Il n'est, en effet, pas nécessaire de comprendre de façon approfondie les différentes règles de gestion contenues dans la définition des processus pour pouvoir adopter notre approche. En effet, cette dernière se limite à déterminer les données d'entrées, de sorties et les ressources utilisées par les différents processus.

Actuellement nous sommes en cours d'application de l'approche pour un nombre plus importants des processus de l'université, incluant la gestion des inscriptions, de la scolarité, etc. D'un point de vue outillage, nous pensons nécessaire de disposer d'un référentiel des processus déployés en utilisant les données liées et du web sémantique (Zemmouchi-Ghomari, 2015). La construction d'un tel référentiel permettra alors de s'affranchir des interviews et d'augmenter le taux d'automatisation de l'approche d'évaluation. D'un autre côté, ce référentiel permettra une meilleure gestion de l'évolution des processus et l'analyse de l'impact de cette évolution sur les performances.

Bibliographie

Badillo P.-Y. (1999). *La méthode dea: Analyse des performances*. Hermes Science.

Bouneffa M., Ahmad A. (2013). The change impact analysis in bpm based software applications: A graph rewriting and ontology based approach. In *Enterprise information systems*, p. 280–295. Springer.

Bouneffa M., Ahmad A., Basson H. (2016). Gestion intégrée du changement des modèles de processus métier. In *Actes du xxxivème congrès inforsid, grenoble, france, may 31 - june 3, 2016.*, p. 33–48.

Cavaignac L. F., Villesèque-Dubus F. (2014). L'apport de la méthode dea au pilotage de la performance des centres de coût : l'exemple de la logistique amont. *Finance Contrôle Stratégie [En ligne]*, vol. 17, n° 3.

Charnes A., Cooper W., Rhodes E. (1981). Data envelopment analysis : Approach for evaluating program and managerial efficiency with an application to the programm follow through experiment in us public school education. *Management Science*, vol. 27, n° 6.

Diagne D. (2006). Mesure de l'efficience technique dans le secteur de l'éducation: une application de la méthode dea. *Swiss Journal of Economics and Statistics (SJES)*, vol. 142, n° II, p. 231–262.

Doux G., Jouault F., Bézivin J. (2009). Transforming bpmn process models to bpel process definitions with atl. In *5th international workshop on graph-based tools.*

Fellmann M., Bittmann S., Karhof A., Stolze C., Thomas O. (2013). Do we need a standard for epc modelling? the state of syntactic, semantic and pragmatic quality. In *International workshop on enterprise modelling and information systems architectures (emisa 2013).* St. Gallen.

Glasson B. C., Hawryszkiewycz I., Underwood A., Weber R. (Eds.). (1994). *Business process re-engineering: Information systems opportunities and challenges* (vol. A-54). Elsevier.

Haworth D. A., Pietron L. R. (2006). Sarbanes-oxley: Achieving compliance by starting with iso 17799. *IS Management*, vol. 23, nº 1, p. 73-87. Consulté sur http://dblp.uni-trier.de/db/journals/ism/ism23.html#HaworthP06

HENDI H. I., BOUNEFFA M., AHMAD A., FONLUPT C. (2016). Ontology based engine for solving passenger train optimization problem. In *Proceedings of aic-mitc 2016. to appear.* IEEE Computational Intelligence Society.

LA VILLARMOIS O. de. (1999). Evaluer la performance des rÉseaux bancaires : La méthode dea. *Décisions Marketing*, nº 16, p. 39-51. Consulté sur http://www.jstor.org/stable/40592663

Mazanek S., Hanus M. (2011). Constructing a bidirectional transformation between bpmn and bpel with a functional logic programming language. *Journal of Visual Languages & Computing*, vol. 22, nº 1, p. 66 - 89. (Special Issue on Visual Languages and Logic)

Newcomer E., Lomow G. (2005). *Understanding soa with web services.* Addison-Wesley.

OMG. (2011). *Business process model and notation (bpmn) version 2.0.* (OMG Document Number: formal/2011-01-03)

Ouyang C., Aalst W. M. van der, Dumas M., Hofstede A. H. ter. (2009, August). From business process models to process-oriented software. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, vol. 19, nº 1, p. 1–37.

Palmer N. (2009). Xml process definition language. In L. LIU, M. T. ÖZSU (Eds.), *Encyclopedia of database systems*, p. 3601–3601. Boston, MA, Springer US. Consulté sur http://dx.doi.org/10.1007/978-0-387-39940-9_1550

Prud'hommeaux E., Seaborne A. (2008, janvier). *Sparql query language for rdf*. W3C Recommendation. W3C. Consulté sur http://www.w3.org/TR/rdf-sparql-query/

Recker J. C., Mendling J. (2006). *On the translation between bpmn and bpel: Conceptual mismatch between process modeling languages.* Namur University Press.

Seiford L. M. (1997). A bibliography for data envelopment analysis (1978-1996). *Annals of Operations Research*, vol. 73, nº 0, p. 393–438. Consulté sur http://dx.doi.org/10.1023/A:1018949800069

Wei Q. (2001). Data envelopment analysis. *Chinese Science Bulletin*, vol. 46, nº 16, p. 1321–1332. Consulté sur http://dx.doi.org/10.1007/BF03183382

Zemmouchi-Ghomari L. (2015, octobre). Linked data, towards realizing the web of data: An overview. *Int. J. Technol. Diffus.*, vol. 6, nº 4, p. 20–39. Consulté sur http://dx.doi.org/10.4018/IJTD.2015100102

Annexe 1

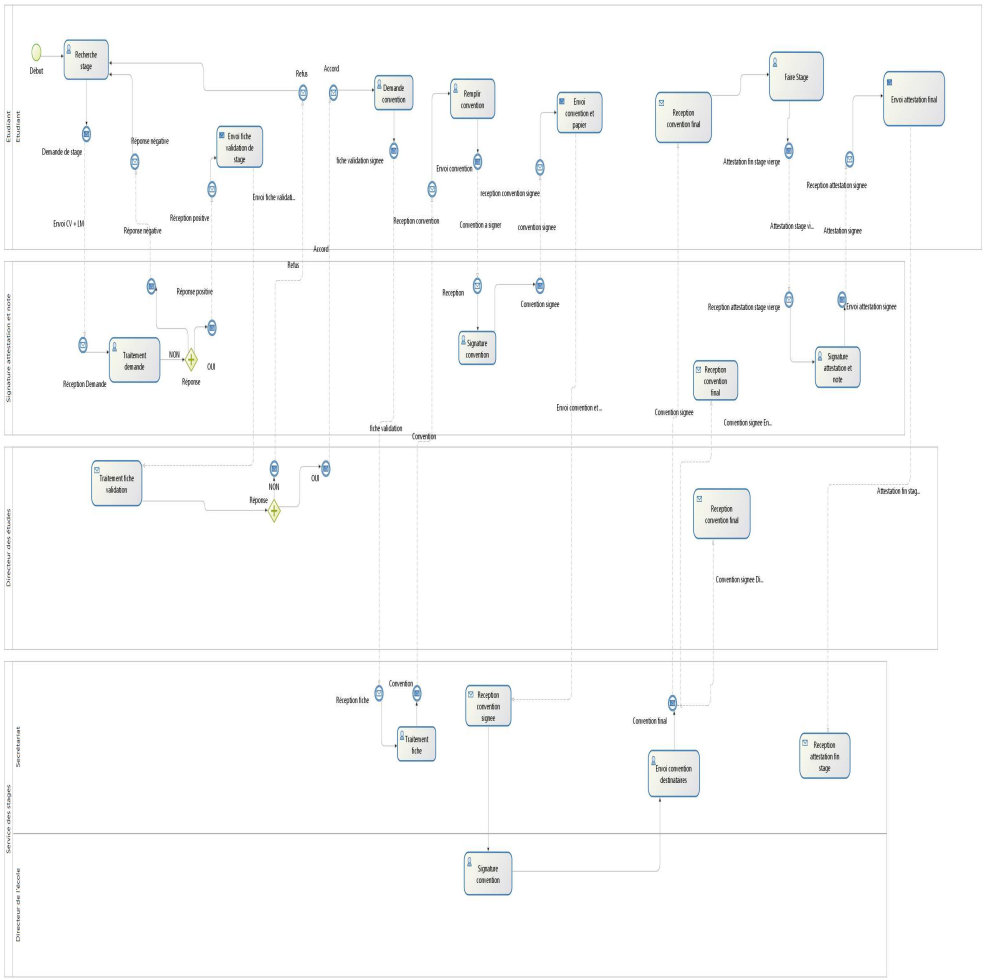


FIGURE *5. Diagramme BPMN du processus de gestion des stages de l'école d'ingénieurs*

Patrons de conception

Patrons temporels pour spécifier les systèmes auto-adaptatifs

YAHIAOUI Ayoub[1], BENJENNA Hakim[1], ROOSE Philippe[2]

1. Laboratoire LAMIS, Université Larbi Tébéssi,
Route de Constantine, 12002 Tébessa, Algérie
ayoub.yahiaoui@univ-tebessa.dz , hakim.bendjenna@univ-tebessa.dz

2. Laboratoire LIUPPA, Univ Pau & Pays Adour,
EA 3000, 64600 Anglet, France
Philippe.Roose@iutbayonne.univ-pau.fr

RESUME : La demande est croissante de systèmes qui nécessitent une adaptation. Ces systèmes doivent avoir la capacité d'adapter leur comportement de façon autonome durant l'exécution en fonction de l'évolution de leur environnement. Parmi les applications nécessitant une capacité d'auto-adaptation : les systèmes automobiles, de télécommunication, de surveillance et les systèmes de maison intelligente. Cependant, malgré son importance, l'auto-adaptation est souvent construite de manière ad-hoc. Dans cet article, nous présentons « Pattern-based Specification for Self-Adaptive Systems (PSAS) », un langage de spécification pour l'auto-adaptation avec un outil support pour faciliter le processus de spécification. La sémantique est présentée en termes de logique floue. Ainsi, un traitement minutieux des exigences, afin de formuler avec précision des exigences relatives aux systèmes auto-adaptatifs, facilite la conception de systèmes flexibles et adaptatifs de manière systématique. Pour montrer l'applicabilité et l'efficacité de notre langage, nous l'appliquons sur un protocole de communication dans les réseaux de capteurs.

ABSTRACT. There is a growing demand for systems that require adaptation. These systems must have the ability to adapt their behavior independently during execution according to the evolution of their environment. Among applications requiring a self-adaptive capacity: automotive systems, telecommunication systems, monitoring systems and intelligent home systems. However, despite its importance, self-adaptation is often constructed in ad-hoc manner. In this paper, we present "Pattern-based Specification for Self-Adaptive Systems (PSAS)", a specification language for self-adaptation with a support tool to facilitate the specification process. Semantics is presented in terms of fuzzy logic. Thus, careful processing of requirements, in order to accurately formulate requirements for self-adaptive systems, facilitates the design of flexible and adaptive systems in a systematic manner. To demonstrate the applicability and effectiveness of our language, we apply it to a protocol of sensors network.

MOTS-CLES Ingénierie des exigences, Systèmes auto-adaptatifs, Patron de spécification, Logique Temporelle Métrique Floue.

KEYWORDS Requirement engineering, Self-adaptive systems, Specification patterns, Fuzzy metric temporal logic.

1. Introduction

Au fur et à mesure que les applications deviennent de plus en plus volumineuses, encore plus hétérogènes et complexes, il existe certaines tâches où les données contextuelles ne peuvent pas être définies durant la phase de conception comme par exemple dans un système de bureau intelligent où les appareils doivent être synchronisés (exigence), mais le nombre des appareils ne peut pas être connu durant la conception. A la lumière de cela, il devient vital pour ces systèmes de s'adapter automatiquement aux changements qui se produisent dans l'environnement comme le nombre de périphériques dans l'exemple précédent. Nous appelons ces systèmes des "Systèmes Auto-adaptatif (SAA)". L'auto-adaptation présente une approche prometteuse pour la gestion de la complexité des systèmes actuels (Iftikhar et Weyns, 2012). Les applications nécessitant des capacités de type SAA sont par exemple des systèmes d'infrastructure intelligents, les réseaux de capteurs et les systèmes embarqués. Les changements de facteurs de l'environnement comme les interactions humaines (entrées imprévues) rendent l'analyse difficile de tous les états dans lesquels le système sera pendant sa durée de vie. Par conséquent, un SAA doit être capable de s'adapter à un ensemble de contextes environnementaux, mais la nature exacte de ces contextes reste vaguement comprise. Un défi global dans le développement des SAAs est la façon d'exprimer une spécification afin de rendre le SAA capable de gérer les problèmes posés par les domaines d'application, y compris les incertitudes comportementales (Whittle et al., 2009). Pour clarifier le concept de SAA, nous rappelons quelques définitions :

- **Définition 1**. Un système auto-adaptatif évalue son propre comportement et modifie sa propre performance lorsque l'évaluation indique que n'est pas accompli ce que le logiciel est censé faire, ou lorsque de meilleures fonctionnalités ou performances sont possibles (Salehie et Tahvildari, 2005).
- **Définition 2**. Un système auto-adaptatif est un système en boucle fermée qui peut se modifier lui-même dû aux changements continus du système, ses exigences et les tendances existantes de développement et de déploiement des systèmes complexes, réduisant les interactions humaines. La conception de systèmes auto-adaptatifs dépend des besoins de l'utilisateur et des propriétés et caractéristiques environnementales du système. Les logiciels auto-adaptatifs nécessitent une grande fiabilité, robustesse, adaptabilité et disponibilité (Weyns et al., 2012).

Les SAAs ont besoin de spécifications flexibles tout en étant formelles, d'où la nécessité d'une logique souple. Nous choisissons la logique floue (Zadeh, 1965) afin de gérer des situations où l'environnement contient des incertitudes.

Dans de nombreux systèmes réels avec incertitude, l'application de la logique floue, a donné des résultats très fructueux. Elle est le résultat de plus de 3 décennies d'études dans le domaine de la spécification et de la vérification et offre de nombreux avantages aux praticiens. Elle reste néanmoins difficile (Dwyer et al., 1999) en raison de la distance importante entre les formalismes employés par les outils de vérification

des modèles et le langage naturel des exigences (Abid et al., 2011). Cette limitation implique l'expression de propriétés à un niveau d'abstraction élevé en utilisant un langage basé sur l'anglais structuré (Konrad et Cheng, 2005), qui est naturel et limité par le formalisme.

Cet article présente un nouveau langage de spécifications basé sur des patrons pour les SAAs afin de faciliter l'expression des contraintes temporelles d'auto-adaptation.

La suite de l'article est organisée comme suit : dans la section suivante, nous définissons les concepts nécessaires pour notre catalogue de patrons. La section 3 se concentre sur nos patrons et leurs sémantiques, nous présentons dans cette section la grammaire d'anglais structuré pour notre langage. Dans la section 4, nous présentons notre outil de support ainsi que nous justifions l'efficacité de notre langage via des instances du monde réel, nous renforçons également notre proposition par une évaluation empirique dans la section 5. Avant de conclure, nous décrivons les travaux connexes effectués dans ce domaine dans la section 6.

2. Contexte

2.1. Patrons de spécification

Un patron de spécification est un modèle pour représenter un sous ensemble de propriétés qui s'expriment de la même manière (Dwyer et al., 1999). Selon (Autili et al., 2015), les patrons sont regroupés en trois familles : les patrons qualitatifs qui décrivent la matérialisation des événements, les patrons temps réel qui étendent la première famille en ajoutant une contrainte temporelle et les patrons probabilistes qui étendent également les patrons qualitatifs en ajoutant la contrainte probabiliste.

2.2. Portée

Afin de décrire le moment où le patron s'applique, chaque classe décrite ci-dessus doit être couplée avec une portée. En d'autres termes, tous les patrons doivent correspondre à la règle "spécification = portée, patron". Le tableau 1 donne les définitions des portées présentées dans (Dwyer et al., 1999). La figure 1 illustre un aperçu du fonctionnement des portées.

Tableau 1 Portée du patron

Portée	Définition
Globally	Le patron peut se maintenir en tout point de l'exécution du système.
After{P}	Le patron s'applique après la première occurrence de P.
Before{P}	Le patron s'applique jusqu'à la première occurrence de P.
Between {P} and {R}	Le patron ne peut s'appliquer qu'entre P et R, autrement dit, si le patron se produit, il doit être précédé de P et suivi de R.

After {P} until {R}	Le patron ne s'applique qu'après l'occurrence de P jusqu'à ce que R soit vrai (R n'est pas nécessaire).

Note : la différence entre "After {P} until {R}" et "Between {P} and {R}" est : dans la première portée, si R ne se produit pas, le modèle est toujours valide. Alors que dans la seconde, il n'est pas.

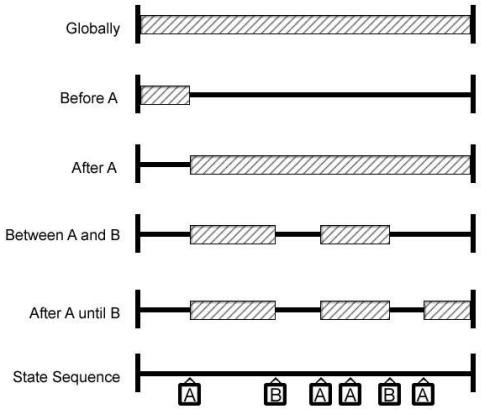


Figure 1 Portée du motif : présente la chronologie autorisée d'un motif spécifique, délimitée par les événements A et B (Dwyer et al., 1999)

3. Présentation du langage

Dans cette section, nous présentons notre langage qui permet aux ingénieurs d'exprimer les propriétés du système de manière flexible. Nous avons conçu des patrons afin de permettre aux concepteurs de savoir que l'exigence peut être changée durant l'exécution lorsqu'un changement environnemental se produit. Le système doit être capable d'ignorer temporairement les exigences non critiques afin de s'assurer que les exigences essentielles puissent être satisfaites. Notre langage supporte la spécification de sources multiples d'incertitude (Whittle et al., 2009 ; Esfahani et Malek, 2013), de manière déclarative flexible (as early as possible) ou d'énumérer les alternatives possibles (may R1 or may R2…) ce qui permet au développeur d'introduire l'auto-adaptation dans le système, par exemple : « le message d'acquittement doit parvenir dans 10 millisecondes », cette exigence est invalide si le message parvient dans 12 millisecondes. Par contre si on utilise le patron « TheEarliestPossible » (présenté ci-dessous) l'exigence est toujours valide mais avec un degré de vérité inferieur à 1.

Notre langage présente plusieurs atouts. Le premier est la possibilité que les exigences soient soulagées par la souplesse de l'expression du patron (as possible), qui se traduit après par la logique floue, afin d'assurer le fonctionnement régulier du système aussi longtemps que possible. Le second atout qu'il repose sur la grammaire anglaise structurée présentée dans (Konrad et Cheng, 2005 ; Autili et al., 2015), ce qui fait de lui un langage sans ambiguïté.

3.1. Catalogue de patron auto-adaptatif

Les patrons (abordés dans la section 2) sont organisés en deux catégories : l'occurrence et l'ordre. La première catégorie présente la matérialisation de l'événement tandis que la seconde capture une séquence d'événements. Nous proposons dans ce papier un catalogue de patrons auto-adaptatifs. Pour des raisons de place, nous ne proposons ici qu'un sous-ensemble de patrons de spécification auto-adaptatifs. Ils peuvent être considérés comme des versions soulagées de ceux existants, par exemple : "TheLongestPossible" peut être considéré comme une version soulagée du patron "Universality" (Dwyer et al., 1999). La classification des patrons proposés est la même que celle présente dans (Autili et al., 2015). Les patrons sont divisés en deux classes principales, l'occurrence auto-adaptative et l'ordre auto-adaptatif.

3.1.1. Patrons auto-adaptatifs d'occurrence

Utilisés pour affirmer de manière flexible qu'une certaine configuration de propriétés doit toujours, éventuellement avoir lieu ou ne doit pas se produire.

- TheLongestPossible : nous proposons ce patron afin de gérer des situations où la propriété doit être vraie le plus longtemps possible.
- TheShortestPossible : ce patron est à l'opposé du premier, il décrit une propriété devant être vraie sur une période la plus courte possible.
- TheEarliestPossible : ce patron a pour objectif que la propriété se produise le plus tôt possible.
- TheLatestPossible : contrairement, ce patron exprime la propriété qui se produit le plus tard possible.
- SA-MinDuration : ce patron implique que la propriété est vraie pendant un temps minimum assoupli pour permettre à la propriété de tenir en dessous.
- SA-MaxDuration : ce patron indique que la propriété doit tenir sous un seuil de temps, le seuil est assoupli pour permettre à la propriété de tenir au-dessus.

3.1.2. Patrons auto-adaptatifs d'ordre

Ils sont utilisés pour spécifier l'ordre dans lequel certaines propriétés doivent se produire. Dans cet article, nous proposons une version auto-adaptative pour les deux patrons "Until" et "Response".

- SA-Until : le patron "Until" est natif dans la plupart des logiques temporelles. Il a été étendu dans (Grunske, 2008), afin de gérer les propriétés temporisées. Nous proposons une version soulagée de ce dernier avec la cartographie liée à chaque portée, où le temps associé est flou.
- TheEarliestPossible-Response : ce patron indique qu'à chaque fois qu'une propriété X est vraie, elle doit être suivie par une autre propriété Y, aussitôt que possible (seule la première occurrence de la seconde propriété est considérée).

Le catalogue de patrons présenté ci-dessus vise à apporter plus d'expressivité durant la phase de spécifications. Il est généré à partir de la grammaire présentée dans ce qui suit, sa sémantique est également présentée dans la sous-section 3.3.

3.2. Syntaxe du langage

La formulation des spécifications est soumise à la grammaire anglaise structurée présentée ci-dessous. Dans la première étape du processus de spécification, chaque propriété est représenté par le couple "portée, patron", ensuite on choisit la portée (présenté dans la section 2). En ce qui concerne la partie du patron (pattern), elle est générée par le non-terminal "Self-adaptative-Pattern". Ce dernier génère deux non-terminaux, Self-adaptive-Occurrence ou Self-adaptive-Order, ensuite le modèle non-terminal du patron, et enfin l'anglais structuré, le tableau 2 montre la grammaire d'anglais structurée utilisée dans ce papier.

Tableau 2 Grammaire structurée du langage A,B,C ∈{Exigences}, tl^A ,$tu^A \in R^+$, tl^A,tu^A sont les bornes d'occurrence inférieure et supérieure, respectivement.

Property	::=	Scope, Self-adaptive-Pattern
Scope	::=	**Globally** | **Before** *{B}* | **After** *{C}* | **Between** *{C}* **and** *{B}* | **After** *{C}* **until** *{B}*
Self-adaptive-Pattern	::=	Self-adaptive-Occurrence | Self-adaptive-Order
Self-adaptive-Occurrence	::=	TheLongestPossible | TheShortestPossible | TheEarliestPossible | TheLatestPossible | SA-MinDuration | SA-MaxDuration
Self-adaptive-Order	::=	SA-Until | TheEarliestPossibleResponse
TheLongestPossible	::=	**It is the case that** *{***A***}* **[holds] as long as possible** [Time (A)].
TheShortestPossible	::=	**It is the case that** *{***A***}* **[holds] as short as possible** [Time (A)].
TheEarliestPossible	::=	**It is the case that** *{***A***}* **[holds] as early as possible** [Time (A)].
TheLatestPossible	::=	**It is the case that** *{***A***}* **[holds] as late as possible** [Time (A)].
SA-MinDuration	::=	**Once** *{***A***}* **[becomes satisfied] it remains as possible up to** t_u^A TimeUnits.
SA-MaxDuration	::=	**Once** *{***A***}* **[becomes satisfied] it remains as possible less than** t_u^A TimeUnits.
SA-Until	::=	**{A} [holds] without interruption until {D} [holds]** [Time(A)].

TheEarliestPossibleResponse	::=	**if {A} [has occurred] then in response {D} [holds] as early as possible** [Time(D)].
Time(A)	::=	UpperTimeBound(A) | LowerTimeBound(A) | Interval(A)
UpperTimeBound(A)	::=	**within** t_u^A TimeUnits
LowerTimeBound(A)	::=	**after** t_l^A TimeUnits
Interval(A)	::=	**between** t_l^A **and** t_u^A TimeUnits
TimeUnits	::=	**any denomination of time (e.g., seconds, minutes, hours, days, or years)**
QuantityUnits	::=	**any denomination of quantity (e.g., number of connected devices)**

3.3. Sémantique du langage

La sémantique de PSAS est définie en termes de Logique Métrique Temporelle Floue (LMTF) (Zhou, 1999). LMTF peut décrire un réseau de pétri temporel avec des informations temporelles incertaines. C'est la représentation de l'incertitude dans LMTF qui la rend approprié comme formalisme pour notre langage. Par exemple, la déclaration "**After**{B},{A} **holds as early as possible**", que "A" exprime une exigence qui survient après l'apparition de l'événement "B", mais il est incertain combien de temps il faut pour que "A" se produit après l'événement "B". La déclaration exprime simplement le désir de la période après l'occurrence de "B" soit aussi petite que possible. Une logique avec une incertitude intégrée est donc nécessaire pour formaliser la sémantique de PSAS.

Un ensemble flou est un ensemble dont les éléments ont des degrés d'appartenance. Dans la théorie des ensembles classiques, un membre appartient à un ensemble ou non. La théorie des ensembles flous permet l'évaluation progressive de l'appartenance à des éléments dans un ensemble, qui est décrit en utilisant une fonction d'appartenance dans la plage des nombres réels [0,1]. En d'autres termes, un ensemble flou est une paire (A, m) où A est un ensemble et m: A→[0,1]. Un nombre flou est un sous-ensemble flou de nombres réels dont la fonction d'appartenance est convexe et normalisée, c'est-à-dire max (m(a) = 1). Un nombre flou peut être triangulaire ou trapézoïdale, dans le sens où son graphique d'appartenance décrit un triangle ou un trapèze avec un Vertex montrant l'appartenance à 1. Par exemple, un nombre flou "2", la valeur précise du nombre est incertaine, en d'autres termes, le nombre représente environ 2. La fonction d'appartenance trapézoïdale indique que toute valeur inférieure à 1,5 ou supérieure à 2,5 n'est certainement pas considérée comme approximativement 2, que [1.75, 2.25] est absolument considéré comme étant environ 2, alors que les valeurs comprises entre 1.5 et 2.5 sont environ 2 avec des degrés de vérité différents. La notion de nombre flou est facilement étendue à une durée floue. La durée $d \epsilon R^+$ est une durée floue s'il existe une incertitude floue sur la durée exacte de la durée. C'est-à-dire qu'il est associé à un nombre flou qui définit une longueur de temps floue.

□ est l'opérateur "toujours" habituel. **U** désigne "jusqu'à" comme avec la logique temporelle standard. **O**, qui prend la valeur de vérité de sa formule après une durée. La durée $d \in R^+$ peut être une durée floue ou non. L'expression $\mathbf{O}_{=d}$ signifie "après exactement d", $\mathbf{O}_{<d}$ représente "avant que *d* soit passé", Et $\mathbf{O}_{>d}$ est "après que *d* soit passé". Par conséquent, si *d* est flou, l'opérateur de délai peut être utilisé pour exprimer des relations avec un intervalle de temps incertain. Les notations abrégées habituellement utilisées sont également disponibles. En particulier, ◊ A = vraie **U** A, où ◊ signifie finalement.

Nous sommes maintenant prêts à définir la sémantique des expressions PSAS en termes de LTMF. Les définitions pour ces opérateurs, y compris l'incertitude, dépendent d'une durée flou ou d'un ensemble flou. Ceux-ci ont généralement un maximum à un point particulier, puis se replient progressivement à l'infini, c'est-à-dire qu'ils ont un graphique d'appartenance trapézoïdale qui est asymptotique. Par exemple, dans le cas de "TheEarliestPossible", la fonction d'adhésion à son maximum à l'instant actuel. Cependant, le patron "TheEarliestPossible" techniquement permet de devenir vrai à tout moment après l'instant actuel. Par conséquent, la fonction d'appartenance pour la durée n'est jamais nulle mais approche de zéro progressivement à mesure que le temps augmente.

Nos patrons sont des versions étendues, en majeure partie, de ceux présentés dans (Autili et al., 2015), ces patrons sont une version relâchée de ceux qui existent, la cartographie est étendue à partir de (Dwyer et al., 1999 ; Konrad et Cheng, 2005 ; Autili et al., 2015) avec des contraintes temporelles floues. Pour les patrons comme «**TheLongestPossible »** ou «**TheEarliestPossible**», il n'y a pas de cartographie existante. Nous proposons une nouvelle cartographie pour eux en FMTL. Ci-dessous, nous montrons la cartographie du patron d'occurrence **TheLongestPossible**. La cartographie des autres patrons se trouve sur le site http://www.psas-tool.com/support-tool/sa-patterns.

Globally	$\Box_{[\![time\_A]\!]}\left(A \wedge \Diamond_{\leq [\![end\_A]\!]} \neg A\right)$
Before B	$\Diamond_{\geq [\![tl\_A]\!]} B \rightarrow \left(\left(A \wedge \Diamond_{\leq [\![end\_A]\!]} \neg A\right) \mathrm{U}_{[\![time\_A]\!]}\ B \vee \Box_{[\![time\_A]\!]}\left(A \wedge \Diamond_{\leq [\![end\_A]\!]} \neg A\right)\right)$
After C	$\Box\left(C \rightarrow \Box_{[\![time\_A]\!]}\left(A \wedge \Diamond_{\leq [\![end\_A]\!]} \neg A\right)\right)$
Between C and B	$\Box\ \Big(\left(C \wedge \Box_{\leq [\![tl\_A]\!]} \neg B\right) \wedge \Diamond_{\geq [\![tl\_A]\!]} B \rightarrow \left(\left(A \wedge \Diamond_{\leq [\![end\_A]\!]} \neg A\right) \mathrm{U}_{[\![time\_A]\!]}\ B \vee \Box_{[\![time\_A]\!]}\left(A \wedge \Diamond_{\leq [\![end\_A]\!]} \neg A\right)\right)\Big)$
After C until B	$\Box\left(\left(C \wedge \Box_{\leq [\![tl\_A]\!]} \neg B\right) \rightarrow \left(\left(A \wedge \Diamond_{\leq [\![end\_A]\!]} \neg A\right) \mathrm{W}_{[\![time\_A]\!]}\ B\right)\right)$
Where : $[\![time\_A]\!] = [t_l^A, t_u^A]$, $[\![end\_A]\!] = \infty$, $[\![tl\_A]\!] = t_l^A$	

Dans cette section, nous avons présenté notre catalogue de patrons, pour traiter l'expression de l'auto-adaptation dans la phase de spécification des exigences. Nous avons également spécifié notre grammaire basée sur l'anglais structuré, de la sélection jusqu'au détail du patron. Ensuite, la sémantique du langage a été présentée en termes de logique floue. Cependant, l'efficacité de notre langage doit être testée à travers des instances réelles et des cas d'études.

4. PSAS-tool

Cette section est dédiée à notre outil de support ainsi qu'un ensemble d'exemples de différents systèmes dont les spécifications doivent être flexibles pour assurer l'auto-adaptation.

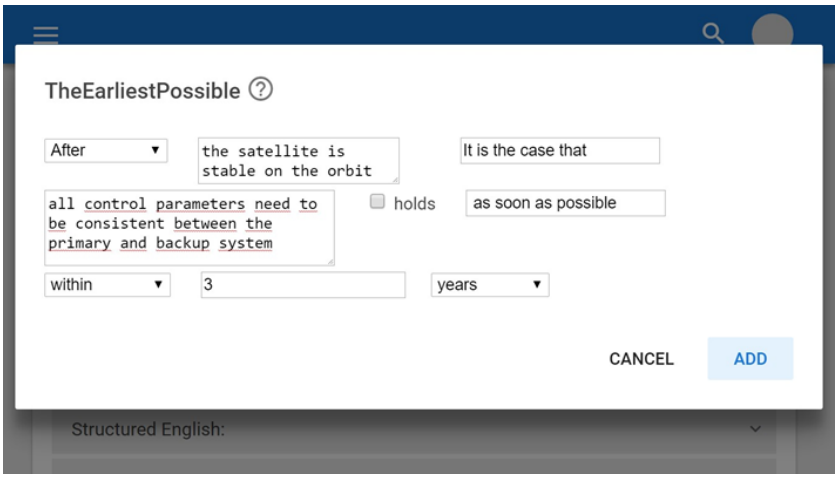


Figure 2 Sélection du patron et création de spécification

4.1. Présentation de PSAS-tool

Pour les concepteurs d'exigences, non familiers avec les logiques temporelles, la spécification des propriétés peut être une tâche difficile (Dwyer et al., 1999). Nous avons donc conçu un outil de support afin de faciliter le travail du concepteur. L'objectif principal de l'outil PSAS-tool est d'automatiser la traduction des expressions de grammaire anglaise structurée en formules logiques temporelles métriques floues (FMTL). Le processus de "PSAS-tool" est comme suit : lors de la sélection du patron dans le menu, une fenêtre modale apparaît (figure 2), puis le remplissage des zones de texte avec l'exigence originale. Après avoir appuyé sur le bouton "ADD", la spécification est insérée dans le document principal et la traduction en FMTL est généré automatiquement (figure 3). Enfin, la liste des spécifications peut être exportée au format pdf.

L'exemple dans la figure 2 montre le patron "TheEarliestPossible":

- Liste des extensions. Elles incluent des événements comme "After{C}", dans ce cas PSAS-tool ajoute automatiquement une zone de texte pour insérer l'événement.

- Zone de texte pour l'événement.

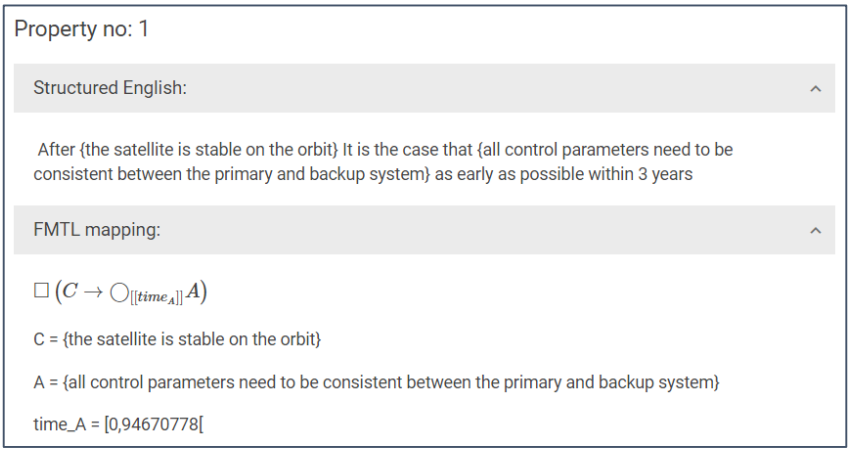


Figure 3 Spécification en anglais structuré et en formule FMTL

- Zone de texte qui décrit l'exigence originale.
- Checkbox optionnelle "holds", pour exprimer clairement l'exigence.
- Définition d'intervalle de temps (within, at least).
- Entrée de la durée liée à la spécification.
- Unité de temps (second, minute, hour ...)

La conception de PSAS-tool permet aux concepteurs des exigences de spécifier les propriétés du système de manière flexible, sans utiliser des symboles logiques. PSAS-tool est une application orienté web, ce choix implique divers avantages, d'abord, les applications web permet d'éviter le déploiement sur chaque machine et facilite les mises à jour, deuxièmement, l'accessibilité de n'importe quel endroit avec accès internet, troisièmement, les applications Web sont indépendantes de la plate-forme. L'outil PSAS-tool fournit également l'avantage de l'adaptabilité mobile grâce à sa conception responsive.

Les spécifications présentées dans la sous-section suivante sont générées par l'application de notre approche sur un ensemble d'instances du monde réel.

4.2 Exemples de patrons auto-adaptatifs

Toutes les spécifications doivent être conformes à la règle : "Portée, patron" (section 3). La portée est l'intervalle de temps, dans lequel le patron peut être valide. La deuxième partie représente le corps du patron. Dans le premier exemple ci-dessous, la portée est "Globally" ce qui signifie que le patron est valable pour tout le temps d'exécution du système. Le modèle ici est "TheLongestPossible", il sera par la suite remplacé par "It is the case that {P} holds as long as possible", les parties en gras restent inchangées, tandis que "P" sera remplacé par l'exigence initiale.

TheLongestPossible

Spécification = Portée, Patron

= Globally, TheLongestPossible.

= Globally, It is the case that {P} holds as long as possible [Time(P)]. */Time(P) optionnel, s'il n'est pas spécifié donc Time(P) = [0,∞[.*

= Globally, It is the case that {Terrestrial Photovoltaics (PV) systems face the sun} holds as long as possible.

TheEarliestPossible

Patron: After{C}, It is the case that {A} holds as early as possible [Time(A)].

Exemple: After {the satellite is stable on the orbit}, It is the case that{all control parameters need to be consistent between the primary and backup system for the 3 year mission time} holds as early as possible. (Source: satellite control system).

SA-MaxDuration

Patron: Globally, Once {A} [becomes satisfied] it remains as possible less than tuATimeUnits.

Exemple: Globally, Once {insulin pump starts injection} becomes satisfied it remains as possible less than d/ d= Duration of insulin action (DIA). (The Fault-Tolerant Insulin Pump Therapy (Capozucca, 2006))

Les exemples en dessus sont représentés par sous-ensemble de motifs. Dans la section suivante, nous présentons un cas d'étude concret pour illustrer comment notre langage ajoute de la souplesse aux exigences.

5 Evaluation empirique

L'objectif est ici de montrer que les patrons de spécification présentés traitent des problèmes actuels de spécifications auto-adaptatives. Nos patrons de spécification sont conçus principalement pour l'expression du temps. Nous présentons ici un cas d'étude sur le protocole de communication "Message Queue Telemetry Transport (MQTT)" où le temps intervient. Le choix de MQTT est également approprié par le fait que MQTT est normalisé et adapté au réseau de capteurs, qui est l'un des domaines d'application actifs des techniques d'auto-adaptation (Macias-Escriva, 2013). Nous présentons les principales opérations incluant le processus de connexion, de publication, de souscription et de désabonnement. MQTT est un protocole de transport de messagerie de publication/abonnement client-serveur. Il est idéal pour une

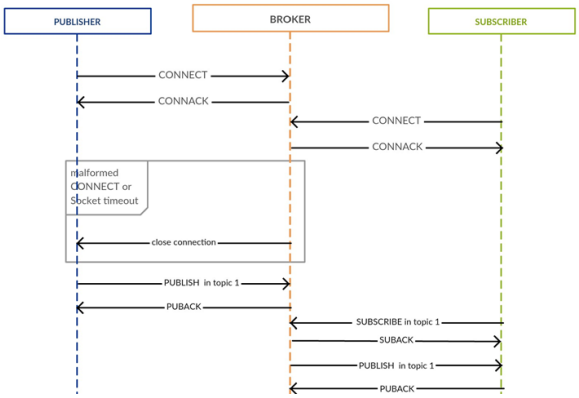


Figure 4 Aperçu sur le protocole MQTT

utilisation dans des environnements avec contraintes comme l'Internet of Things (IoT) et le Machine to Machine (M2M). La figure 4 présente une vue d'ensemble du protocole MQTT dans un diagramme de séquence simplifié.

Dans ce qui suit, nous présentons un ensemble de spécifications MQTT en utilisant les patrons de (Autili et al., 2015) pour les exigences invariantes et notre catalogue proposé pour les exigences d'auto-adaptation.

Exigences invariantes

- Globally, it is never the case that {a client is connected to another directly} holds.
- Globally, it is always the case that {all communications get through the broker} holds.

Processus de connexion (exigences non-invariantes)

TheEarliestPossible-Response: Globally, if {the client sends a CONNECT message to the broker} then in response {the broker sends a CONNACK and a status code to the client} as early as possible.

TheLongestPossible: Before{a client sends a disconnect command or loses connection}, It is the case that {the broker keep the established connection open} holds as long as possible.

SA-max-duration: Before{the broker close the connection}, Once {the broker waits CONNECT message from the client} it remains as possible less than MAXTIME. / MAXTIME = a reasonable amount of time defined by the broker.

TheEarliestPossible-Response: After{a client is connected}, if{a client set CleanSession to true} then in response {the broker sends a CONNACK with session flag is false to the client} as early as possible.

After{client is connected}, if{client has set CleanSession to false and stored client session information exist} then in response {the broker sends to the client a CONNACK with session flag is true} as early as possible.

After{client is connected}, if{client has set CleanSession to false and stored client session information does not exist} then in response {the broker sends to client a CONNACK with session flag is false} as early as possible.

Publier

- After{client is connected}, if{a client sends a publish to the broker} then in response {the broker will read the publish} holds as early as possible followed by({the broker acknowledge the publish if needed (according to the QoS Level)}{the broker process the publish}).

Abonnement/Désabonnement

- After{client is connected}, if{a client sends a SUBSCRIBE/ UNSUBSCRIBE message to the MQTT broker} then in response {the broker sends a SUBACK/UNSUBACK message to client} holds as early as possible.

Les patrons proposés dans les exemples ci-dessus sont plus souples que les patrons existants pour les systèmes non adaptatifs. PSAS à l'avantage par deux aspects principaux. Tout d'abord, les spécifications générées sont simples grâce à sa grammaire non récursive. Deuxièmement, notre langage offre un moyen de gérer l'intervalle d'exécution des spécifications via le concept de portée.

6 Travaux connexes

C. Krupitzer et al. (2015) ont parlé des systèmes auto-adaptatifs et des défis dans ce domaine. Faire face à l'incertitude est l'un des défis importants pour le domaine de l'ingénierie des exigences. Il peut être matérialisé dans un nouveau langage pour spécifier les exigences (Dey, 2001). Un travail similaire à celui proposé dans cet article est RELAX (Whittle et al., 2009). Les auteurs visent, par leur langage, à gérer l'incertitude d'une manière déclarative à l'aide d'opérateurs temporels, ordinaux et modaux. Ils choisissent pour leur cartographie, la logique temporelle de branchement floue (Moon et al., 2004), alors que notre langage est basé sur des patrons, avec une grammaire anglaise structurée (Autili et al., 2015) pour traiter l'ambiguïté dans les spécifications. De plus, notre langage permet l'expression de tous les opérateurs RELAX et plus. De manière similaire à RELAX, notre langage vise également à soutenir des adaptations non prévues. Dans un article ultérieur (Cheng et al., 2009), RELAX a été utilisé avec la modélisation d'objectifs pour spécifier l'incertitude dans d'autres sources. Ils construisent d'abord le modèle de but, puis l'utilisent de manière ascendante pour rechercher les sources d'incertitude qui sont les éléments du domaine/environnement et peuvent compromettre la satisfaction des objectifs.

James F. Allen (Allen, 1983) décrit un système de raisonnement sur les intervalles temporels expressif. Cette approche est utile dans les domaines où l'information temporelle est imprécise et relative, cependant cette dernière ne donne pas d'information précise concernant le degré de vérité des expressions. Baresi et al. (2010) traitent l'incertitude des objectifs via FLAGS. Analogue à RELAX, ils ont pour but d'atteindre l'objectif principal des systèmes adaptatifs au niveau des exigences : atténuation de l'incertitude attachée aux besoins de l'environnement en intégrant l'adaptabilité dans le système dès l'élicitation des exigences. Les exigences

spéciales à FLAGS sont appelées objectifs adaptatifs. Ils permettent les stratégies de prévention qui doivent être réalisées si certains objectifs ne sont pas satisfaits comme prévu. FLAGS traite également une autre source d'incertitude. L'incertitude dans les objectifs eux-mêmes. FLAGS est basé sur des objectifs flous pour lesquels les propriétés ne sont pas complètement connues. La spécification complète n'est pas disponible et de petites violations temporaires sont tolérées. Ainsi, FLAGS se terminent par deux séries d'objectifs : nettes et floues. Il formalise les buts en utilisant la logique temporelle floue pour les langages flous et la logique temporelle linéaire pour les autres. FLAGS s'appuie également sur une grammaire ambigüe, ce qui le rend difficile à gérer par des ingénieurs moins expérimentés (Autili et al., 2015).

Tableau 3 Vue générale sur les langages d'auto-adaptation

Langage	Support d'auto-adaptation	Gestion d'incertitude	Ambiguïté	Basé-patron	Cartographie
RELAX	Oui	Oui	Ambigu	Non	FBTL
FLAGS	Oui	Oui	Ambigu	Non	LTL/FTL
PSAS	Oui	Oui	Non-ambigu	Oui	FMTL

Deux nouvelles catégories d'exigence ont été proposées. Les exigences de sensibilisation (Souza et al., 2011) servent à surveiller les exigences du système et les exigences d'évolution dédiées à la représentation des plans d'adaptation afin de faire face aux changements dans les modèles d'exigences (Souza et al., 2012). A. Manzoor et al. (2015) emploient à la fois des approches déclaratives et basées sur des buts. Ils ont utilisé RELAX (Whittle et al., 2009) pour spécifier les exigences non-fonctionnelles et les notions GORE pour susciter et modéliser les exigences des systèmes auto-adaptatifs. E. Vassev et M. Hinchey (2015) ont proposé une approche pour l'ingénierie des exigences d'autonomie (ARE). Cette approche, basée sur l'ingénierie des exigences orientée objectifs (GORE) et les exigences génériques d'autonomie (GAR), a l'objectif d'éliciter et spécifier les exigences d'autonomie ainsi que la définition des objectifs alternatifs, pour les systèmes auto-adaptatifs. D'autres approches ont été proposées, comme les approches basées sur les objectifs et les agents. Ceux fondés sur des modèles d'objectifs sont par exemple i* (Yu, 1997), KAOS (Dardenne et al., 1993), FLAGS (Baresi et al., 2010). Tropos4AS (Morandini et al., 2017) est une approche récente basée sur les agents, qui repose sur Tropos (Bresciani et al., 2004).

Les travaux décrits ci-dessus ont suscité un intérêt particulier pour l'incertitude qui est considérée comme un principal problème dans le développement des systèmes auto-adaptatifs. Le traitement de ce dernier dans les premières phases de développement est nécessaire. Cependant, certains problèmes dans un langage de spécification doivent être manipulés. Par exemple, la grammaire du langage doit être d'une part non ambigüe, et d'autre part le langage naturel dans la spécification est limité en raison de son ambiguïté. Notre travail traite à la fois le langage naturel et

l'ambiguïté grammaticale, via le catalogue de patrons et la grammaire anglaise structurée. Le tableau 3 résume certaines approches similaires.

7 Conclusion

Cet article vise à présenter un nouveau langage de spécification pour les systèmes auto-adaptatifs. L'objectif principal de notre langage est de faire face aux changements d'exécution, y compris l'incertitude, afin de spécifier le comportement d'un système auto-adaptatif pour répondre aux changements inattendus, qui se produisent dans l'environnement d'exécution. L'absence d'informations suffisantes sur le comportement prévu de l'application peut causer une incertitude comportementale pendant la phase de développement, de sorte qu'il nécessite encore une adaptation au moment de l'exécution. Nous avons introduit un catalogue de patrons pour les exigences non-invariantes. Le langage est basé sur deux piliers majeurs : la logique floue et les patrons de spécification. Le premier a comme objectif de traiter à la fois les incertitudes temporelles et ordinales grâce à la flexibilité de la logique floue alors que le second est destiné à traiter les ambiguïtés du langage naturel à travers l'anglais structuré. Nous avons donné un ensemble d'exemples pour apporter un sens à notre proposition. Nous avons également présenté un cas d'étude industriel à partir d'un protocole dédié au domaine des réseaux de capteurs, l'un des secteurs les plus actifs dans l'auto-adaptation.

8 Références

Abid N. et al. (2011). A Real-Time Specification Patterns Language, https://hal.archives-ouvertes.fr/hal-00593965

Autili M. et al. (2015). Aligning qualitative, real-time, and probabilistic property specification patterns using a structured english grammar. IEEE Transactions on Software Engineering, vol. 41, n° 7, p. 620-638.

Banks A., Gupta R. (2014). Version 3.1.1. 29 October 2014. OASIS Standard. s.f.

Baresi L. et al. (2010). Fuzzy goals for requirements-driven adaptation. Requirements Engineering Conference (RE), 18th IEEE International 2010.

Bellini P. et al. (2009). Expressing and organizing real-time specification patterns via temporal logics. Journal of Systems and Software, vol. 82, n° 2, p. 183-196.

Bresciani P. et al. (2004). Tropos: An agent-oriented software development methodology. Autonomous Agents and Multi-Agent Systems, vol. 8, n° 3, p. 203-236.

Capozucca A. (2006). The fault-tolerant insulin pump therapy. Rigorous Development of Complex Fault-Tolerant Systems, Springer Berlin Heidelberg, p. 59-79.

Cheng B. HC. et al. (2009). A goal-based modeling approach to develop requirements of an adaptive system with environmental uncertainty. Model Driven Engineering Languages and Systems, Springer, p. 468-483.

Dardenne A. et al. (1993). Goal-directed requirements acquisition. Science of computer programming, vol. 20, n° 1-2, p. 3-50.

Dey A K. (2001). Understanding and using context. Personal and ubiquitous computing 2001, vol. 5, n° 1, p. 4-7.

Dwyer M. et al. (1999). Patterns in property specifications for finite-state verification. Software Engineering, Proceedings of the 1999 International Conference, IEEE.

Esfahani N., Malek S. (2013). Uncertainty in self-adaptive software systems. Software Engineering for Self-Adaptive Systems II, Springer Berlin Heidelberg, p 214-238.

Gruhn V., Laue R. (2006). Patterns for timed property specifications. Electronic Notes in Theoretical Computer Science, vol. 153, n° 2, p. 117-133.

Grunske L. (2008). Specification patterns for probabilistic quality properties. in Proceedings of the 30th International Conference on Software Engineering, IEEE, p. 31-40.

Konrad S., Cheng B. HC. (2005). Real-time specification patterns. Proceedings of the 27th international conference on Software engineering, IEEE, p. 372-381.

Koymans R. (1990). Specifying real-time properties with metric temporal logic. Real-time systems, vol. 2, n° 4, p. 255-299.

Krupitzer C. et al. (2015). A survey on engineering approaches for self-adaptive systems. Pervasive and Mobile Computing, vol. 17, p. 184-206.

M. Usman Iftikhar and Danny Weyns, A Case Study on Formal Verification of Self-Adaptive Behaviors in a Decentralized System. Proceedings 11th International Workshop on Foundations of Coordination Languages and Self Adaptation, FOCLASA2012, Newcastle, U.K., September 8, 2012, pages 45-62.

Manzoor A. et al. (2015). Modeling and Verification of Functional and Non-Functional Requirements of Ambient Self-Adaptive Systems. Journal of Systems and Software, vol. 107, p. 50-70.

Macias-Escriva F D. et al. (2013). Self-adaptive systems: A survey of current approaches, research challenges and applications. Expert Systems with Applications, vol. 40, n° 18, p. 7267-7279.

Moon S. et al. (2004). Fuzzy branching temporal logic. Systems, Man, and Cybernetics, Part B: Cybernetics, vol. 34, n° 2, p. 1045-1055.

Morandini M et al. (2017). Engineering requirements for adaptive systems. Requirements Engineering, vol. 22, n° 1, p. 77-103.

Rajeev A. (1991). Techniques for automatic verification of real-time systems. Stanford University.

Salehie M., Tahvildari L. (2005). Autonomic computing: emerging trends and open problems. ACM SIGSOFT Software Engineering, vol. 30, p. 1-7.

Souza V E S. et al. (2011). Awareness requirements for adaptive systems. Proceedings of the 6th international symposium on Software engineering for adaptive and self-managing systems, ACM, p. 60-69.

Souza V E S. et al. (2012). (Requirement) evolution requirements for adaptive systems. Proceedings of the 7th International Symposium on Software Engineering for Adaptive and Self-Managing Systems, IEEE, p. 155-164.

Vassev E., Hinchey M. (2015). Engineering Requirements for Autonomy Features. Software Engineering for Collective Autonomic Systems, Springer, p. 379-403.

Weyns D., Iftikhar M. U., Malek S., Andersson J. (2012). Claims and supporting evidence for self-adaptive systems-A literature study, SEAMS'12, Proceedings of the 7th International Symposium on Software Engineering for Adaptive and Self-Managing Systems, p.89-98.

Whittle, J., Sawyer P., Bencomo N., Cheng B. HC., Bruel J. M. (2009). Relax: Incorporating uncertainty into the specification of self-adaptive systems. In Requirements Engineering Conference, 2009. RE'09. 17th IEEE International. IEEE, p. 79-88.

Yu Eric SK. (1997). Towards modelling and reasoning support for early-phase requirements engineering. Requirements Engineering, Proceedings of the Third IEEE International Symposium, IEEE, p. 226-235.

Zadeh L A. (1965). Fuzzy sets. Information and control, vol. 8, n° 3, p. 338-353.

Zhou Y and Murata T. (1999). Petri net model with fuzzy timing and fuzzy-metric temporal logic. International Journal of Intelligent Systems, vol. 14, n° 8, p. 719-745.

Modélisation et génération de bases de données géographiques imprécises pour les systèmes relationnels

Extension de F-Perceptory et dérivation automatique de modèles

Besma Khalfi [1,2], **Cyril de Runz** [1,3,4], **Sami Faiz** [2], **Herman Akdag** [1]

1. *Laboratoire LIASD, Université Paris 8*
   2 rue de la liberté 93526 Saint -Denis cedex, France
   \{khalfi,derunz,akdag\}@ai.univ-paris8.fr
2. *Laboratoire LTSIRS, Université de Tunis el Manar*
   BP 37, LE BELVEDERE 1002 TUNIS
   khalfi.besma@hotmail.com;sami.faiz@insat.rnu.tn
3. *Laboratoire CReSTIC, Université de Reims Champagne-Ardenne*
   IUT de Reims, Chemin des Rouliers, CS30012 51687 REIMS CEDEX 2
   cyril.de-runz@univ-reims.fr
4. *CNRS, LIP6, UMR7606,*
   4 place Jussieu, 75005 Paris, France
   cyril.de-runz@lip6.fr

RÉSUMÉ. *Les données géospatiales sont imprécises par nature : leur qualification et leur quantification sont soit liées à des approximations, soit à des évaluations subjectives. Il est intéressant de développer des méthodes de conception dédiées. Cela a conduit récemment à la proposition de l'approche F-Perceptory. L'approche n'est cependant appropriée que pour modéliser des espaces homogènes de formes simples ne permettant pas de modéliser des espaces complexes avec des structures composites. Notre proposition étend F-Perceptory et définit un processus de dérivation automatique de modèles permettant de générer la base de données relationnelles.*

ABSTRACT. *Geospatial data are imprecise by nature : their qualification and quantification are either linked to approximations or to subjective assessments. It is important to develop design methods that are dedicated to imprecise spatiotemporal data. This leads to the recent proposal of the F-Perceptory approach. F-Perceptory is appropriate to represent homogeneous spaces*

based on simple shapes and does not model complex spaces having composite structures. The proposal extends F-Perceptory and defines an automatic models mapping process in order to generate the relational database.

MOTS-CLÉS : *F-Perceptory, données floues, modélisation, géométries composites floues.*

KEYWORDS: *F-Perceptory, fuzzy data, modeling, fuzzy composite geometries.*

1. Introduction

Les recherches autour du stockage et de l'intégration des données spatiales constituent un volet qui contribue à la structuration, l'intégrité et la qualité des données recueillies. Par ailleurs, de nombreux projets, menés dans différents contextes allant de l'agronomie (e.g. le projet Observox (Zayrit, Desjardin, 2012), à la géographie (notamment pour l'étude des phénomènes de marge et de frontières (De Ruffray, 2007)) en passant par l'archéologie (e.g. les projets SIGRem/ArcheoChamps (Desjardin *et al.*, 2012)), ont montré l'importance de l'utilisation de représentations gérant l'imperfection des données géographiques.

Cependant, la prise en compte de l'imperfection des données géographiques, particulièrement de l'imprécision, ajoute une réelle complexification. Plusieurs méthodes et outils ont été développés pour répondre aux défis des besoins réels de la géomatique afin de considérer la nature imparfaite (i.e. imprécision, incertitude, incomplétude) des données géographiques et de l'intégrer dans le processus de modélisation, de stockage et d'analyse. Ainsi, des extensions aux modèles conceptuels classiques ont été proposées à l'instar de celles basées sur le modèle Entité-Relation (Zvieli, Chen, 1986 ; Chen, Kerre, 1998), celles sur le modèle objet (Ma, 2005 ; Sicilia, Garcia-Barriocanal, 2006) ou encore sur les modèles spatiotemporels (Pantazis, Donnay, 1998 ; Shu *et al.*, 2003 ; Zoghlami, 2016).

L'approche F-Perceptory dont les principes et les concepts formels ont été définis dans (Zoghlami, 2016), définit un profil UML pour la modélisation conceptuelle de données spatiotemporelles imprécises. Elle modélise les géométries primitives imprécises (point flou, ligne floue et polygone flou) qui présentent des espaces homogènes. L'approche F-Perceptory ne considère pas initialement les formes composites des objets géographiques. Cependant, le concepteur peut avoir besoin de modéliser des formes spatiales plus complexes, basées sur des structures composites.

Dans cet article, nous présentons les géométries composites floues et étudions leur modélisation conceptuelle. Afin de fournir un modèle de bases de données géographiques floues cohérentes, nous proposons un certain nombre de règles de transformation afin de dériver le modèle conceptuel de données en schéma physique de base de données. Les règles de transformation sont intégrées dans un processus automatique de dérivation de modèles. La version améliorée de F-Perceptory a été implémentée

sous forme de prototype de modélisation, intégré dans l'atelier de génie logiciel libre Modelio [1].

L'article est organisé comme suit : la section 2 présente le positionnement, les fondements et les limites de l'approche F-Perceptory. La section 3 présente la solution proposée pour modéliser les objets géographiques flous à géométries composites et discute de leur dérivation en UML. La section 4 présente un exemple d'application à travers le prototype développé. Enfin, nous concluons et proposons de futures recherches.

2. F-Perceptory

2.1. Positionnement

Depuis son introduction en 1965, la théorie des ensembles flous (Zadeh, 1965) a été intégrée dans de nombreuses approches afin de traiter des informations imparfaites, essentiellement pour la quantification de l'imprécision. En appliquant la théorie des ensembles flous, de nombreuses propositions ont étendu le modèle entité-relation (ER) en introduisant le flou pour les entités, les relations, les attributs et les instances (Zvieli, Chen, 1986 ; Chen, Kerre, 1998 ; Ma *et al.*, 2001). D'autres travaux ont porté sur l'extension du modèle orienté objet (Yazici, Akkaya, 2000 ; Shu *et al.*, 2003 ; Ma, 2005 ; Sicilia, Garcia-Barriocanal, 2006) en proposant diverses extensions concernant les classes floues, les règles floues, la fonction d'appartenance, les associations floues entre les classes, etc.

En ce qui concerne les données géographiques, de nombreuses méthodes ont étendu les modèles conceptuels usuels pour modéliser les données spatiotemporelles telles que CONGOO (Pantazis, Donnay, 1996), MADS (Parent *et al.*, 1997) et PERCEPTORY (Bedard, 1999). Cependant, ces méthodes ne prennent pas en compte la modélisation des données géographiques floues.

En considérant les problèmes d'imperfection, certains chercheurs ont essayé de proposer des solutions pour la modélisation des données géographiques floues (Parent *et al.*, 1997 ; Miralles, 2006). L'approche F-Perceptory (Zoghlami, 2016) a été récemment proposée pour la modélisation de données spatiotemporelles floues.

2.2. Fondements théoriques

Les observations de l'espace révèlent la complexité et la variabilité de la réalité géographique. Contrairement aux espaces géographiques bien définis ayant des limites précises, nettes et linéaires, beaucoup d'autres situations mettent en lumière des objets

1. Modelio est un outil de modélisation complet. Il supporte la norme UML 2.4.1, s'appuie intégralement sur le langage Java et dispose d'un système d'extensions permettant d'étendre ses fonctionnalités (www .modeliosoft.com).

géographiques avec des limites plus ou moins nettes et continues. L'hypothèse fondamentale de la modélisation floue porte sur le fait que chaque objet présente un noyau et des bordures. Le noyau est la partie où tous les éléments en son sein appartiennent pleinement au dit objet alors que les bordures forment la partie où l'appartenance à l'objet est partielle et possiblement graduée.

Par conséquent, la question « Comment caractériser un territoire imprécis ? » est liée aux outils de représentation en termes de modèles mathématiques et de modèles informatiques. La théorie des ensembles flous est l'une des solutions mathématiques définies pour représenter l'imprécision. Grâce à elle, il est possible d'exprimer une appartenance partielle d'une valeur à un ensemble. En effet, si E est un ensemble flou et e est un élément de E, la proposition « e est un membre de E » n'est pas nécessairement soit vraie soit fausse. Elle peut être vraie dans une certaine mesure. L'ensemble E est caractérisé par une fonction d'appartenance μ_E prenant ses valeurs dans [0, 1].

Pour stocker un ensemble flou, une solution possible est de le discrétiser de sorte qu'un nombre fini de α-coupes soit pris en considération afin de le représenter. Une α-coupe est l'ensemble des valeurs du domaine (l'ensemble des e) ayant un degré d'appartenance supérieur ou égal à α ($\mu_E(e) \geq \alpha$). La Figure 1 illustre un exemple de discrétisation de l'ensemble E en trois α-coupes.

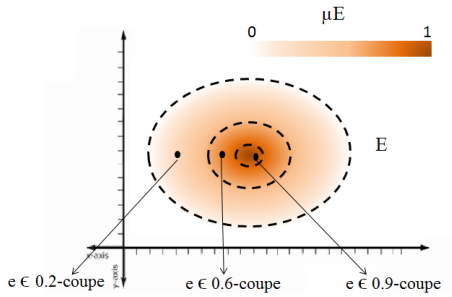


Figure 1. Un ensemble flou E avec trois α-coupes

Une α-coupe de l'ensemble flou E est notée E_α. E_1 est appelée noyau ou cœur de E. E_0, appelée support de E, est l'ensemble des éléments ayant un degré d'appartenance strictement supérieur à 0.

2.3. Langage pictogrammique

L'approche F-Perceptory est basée sur le langage PictograF (Bedard, 1999) et sur les concepts de Fuzzy UML (Ma, 2005) pour définir son langage pictogrammique propre à la modélisation de données spatiotemporelles imprécises. Cette approche considère, en effet, deux types de données imprécises : flou et possibiliste. Pour le cas flou, chaque objet a une définition restrictive avec un degré d'appartenance plus élevé et un certain nombre de définitions moins restrictives avec un degré d'appartenance

inférieur. Pour le cas possibiliste, chaque objet possède un ensemble de définitions possibles avec des valeurs de confiance. Dans le reste de l'article, nous considérons le cas des données floues.

F-Perceptory met en œuvre les principaux concepts du diagramme de classes UML (classes, attributs, associations, généralisation, contraintes et notes). L'approche est utilisée pour définir un diagramme de classes floues. F-Perceptory permet de modéliser l'imprécision liée à la spatialité, la temporalité et celle des attributs quantitatifs non géométriques des objets spatiotemporels à l'aide de représentations floues dédiées :

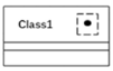
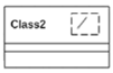
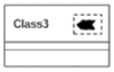
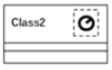
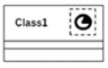
– Pour modéliser la spatialité imprécise, F-Perceptory distingue trois formes spatiales simples : le *point flou*, le *polygone flou* et la *ligne floue*.

– Chaque espace géographique ou phénomène géo-localisé peut avoir une information approximative sur son existence temporelle. F-Perceptory considère l'existence floue instantanée ("*Date floue*") et l'existence floue durable sur un intervalle de temps ("*Période floue*").

– F-Perceptory couvre également l'imprécision sur les valeurs quantitatives (i.e. les attributs dont la valeur appartient à un ensemble flou). Pour ce faire, le nom de l'attribut est précédé par le mot clé "*fuzzy*".

Le tableau 1 présente les pictogrammes F-Perceptory associés à l'imprécision des données géographiques.

Tableau 1. Modélisation pictogrammique de F-Perceptory

Modèle F-Perceptory	**Description**	**Exemple**
	Classe d'objets avec une forme point flou	Centre ville (0 dimension)
	Classe d'objets avec une forme ligne floue	Segment de route ou une rivière (1 dimension)
	Classe d'objets avec une forme polygone flou	Villes, parcs, bâtiments (2 dimension)
	Classe d'objets avec une date floue	Date d'accident (0 dimension)
	Classe d'objets avec une période floue	Période d'existence du patrimoine (1 dimension)

2.4. Dérivation en UML

Pour construire une base de données objet-relationnelles et la remplir avec des données géographiques imprécises, la modélisation conceptuelle doit arriver à une

définition UML correcte du modèle flou. La Figure 2 illustre un exemple de dérivation d'un modèle F-Perceptory en UML. Le modèle F-Perceptory présente une classe d'objets flous à géométrie simple de type polygone flou.

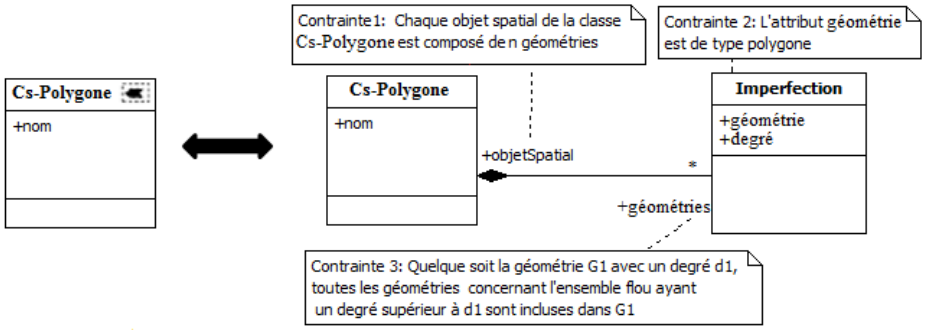


Figure 2. Dérivation UML d'une classe d'objets flous à géometrie polygone flou

Selon (Zoghlami, 2016), chaque objet flou peut être représenté par un ensemble de n α-coupes. La modélisation UML d'une classe d'objets flous génère une nouvelle classe pour la modélisation de l'ensemble des α-coupes appartenant à chaque objet flou. Par conséquent, la classe d'objets flous *Cs-Polygone* est transformée en UML en une classe non stéréotypée ayant une relation de composition avec une nouvelle classe appelée *Imperfection*. La classe *Imperfection* permet de représenter des α-coupes sur un ensemble flou de géométries polygonales. C'est une classe spatiale possédant un attribut *géométrie* renseignant sur la géométrie de chaque α-coupe et un attribut *degré* présentant le degré d'appartenance de l'α-coupe à l'objet *Cs-Polygone*. Dans la relation de composition, le rôle *géométries* référence l'ensemble des α-coupes de chaque objet spatial flou qui lui est le rôle référençant chaque instance de *Cs-Polygone* .

La dérivation en UML engendre la définition de trois contraintes essentielles :

– La première contrainte est liée à la structure de chaque objet spatial flou. Chaque objet *Cs-polygone* doit être composé de n géométries (avec n = nombre des α-coupes).

– La deuxième contrainte se réfère à l'ensemble des α-coupes qui représentent l'objet spatial flou. Les α-coupes doivent former un ensemble flou connexe et normalisé, ce qui signifie que : 1) indépendamment de la géométrie $G1$ de degré $d1$, toutes les géométries autour de l'ensemble flou ayant un degré supérieur à $d1$ sont incluses dans $G1$; 2) les formes sont des formes géométriques connexes ; et 3) le degré d'appartenance maximum est égal à 1.

– La troisième contrainte est liée à la géométrie de chaque α-coupe. Dans le cas d'un objet de type polygone flou, la géométrie de chaque α-coupe doit être de type *polygone*. Dans le cas d'un objet de type point flou, la géométrie de l'α-coupe est de type *point* si le degré d'appartenance est égal à 1, sinon la géométrie de l'α-coupe est de type *polygone*. Dans le cas d'un objet de type ligne floue, la géométrie de l'α-coupe est de type *ligne* si le degré d'appartenance est égal à 1, sinon la géométrie de l'α-coupe est de type *polygone*.

2.5. Limites

L'approche F-Perceptory gère la modélisation des objets géographiques flous à géométries simples. Cependant, selon les contraintes définies sur les propriétés géométriques des objets spatiaux (forme, composition, taille) ou sur l'échelle utilisée (grande ou limitée), nous distinguons divers types de géométries composites dont la modélisation manque dans la définition initiale de F-Perceptory (cf. Figure 3). Ces types décrivent, par exemple, des objets géographiques avec une géométrie facultative (i.e. la géométrie peut-être définie ou pas) ou des objets géographiques dont la géométrie est composée d'une collection de formes de même dimension (e.g. multipoint).

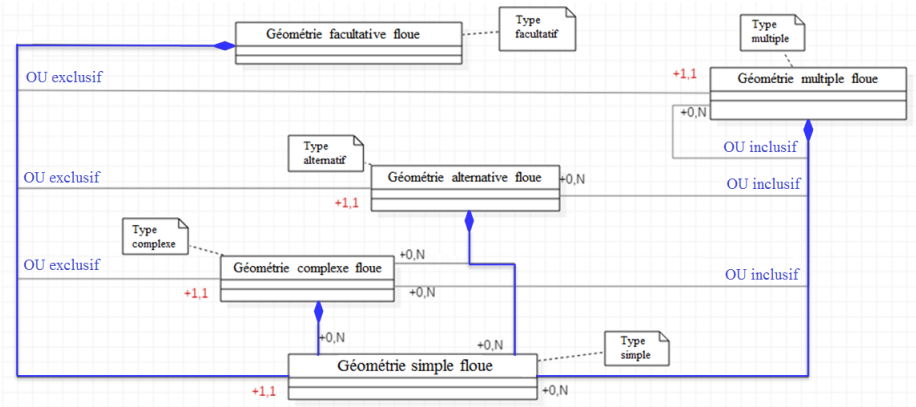


Figure 3. Les types de géométries floues composites pour les objets géographiques : Figure adaptée de (Bedard et al.*, 2002)*

Comme illustré par la Figure 3, les géométries composites floues peuvent être basées sur une composition de formes simples (les relations de composition en bleu) ou composées de formes elles-même composites (les relations de compositions en gris). Dans ce travail, nous considérons le cas des géométries composites basées sur des formes simples.

3. Patrons de conception de géométries composites floues

Nous donnons, à travers des exemples, la présentation pictogrammique des classes d'objets à géométries composites floues et présentons leur transformation en UML. La transformation introduit de nouvelles classes et un certain nombre de contraintes d'intégrité qui sont, dans cette section, exprimées textuellement.

3.1. Patron de conception de géométrie facultative

Implicitement, la multiplicité par défaut des pictogrammes dans les classes est (1, 1). Elle n'est pas exprimée visuellement. Il est supposé que toutes les instances d'une

classe d'objets géographiques ont des propriétés géométriques (forme et cordonnées). Le modèle « *facultatif* » est utilisé lorsque la forme de certaines instances selon certains critères est facultative (Bedard, 1999), comme par exemple, lorsqu'il n'est pas possible d'obtenir la forme géométrique de toutes les instances. Par conséquent, nous mettons la multiplicité (0, 1) à côté du pictogramme pour indiquer que certaines instances peuvent être stockées sans propriétés géométriques.

Considérons, par exemple, la classe d'objets flous *Ville* (cf. Figure 4). En fonction des données disponibles, certaines villes seront stockées dans la base de données sans propriétés géométriques tandis que d'autres le seront par l'intermédiaire d'un polygone flou.

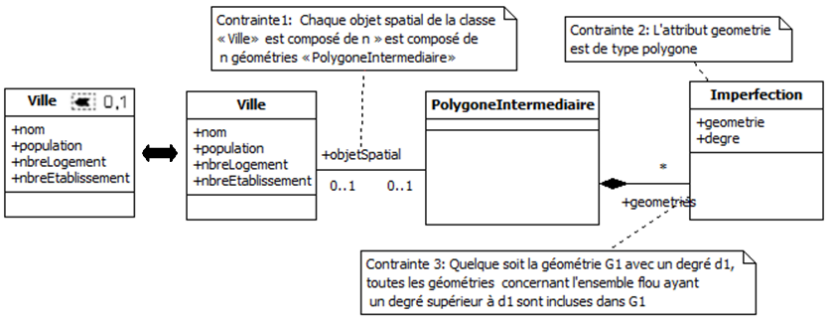


Figure 4. Dérivation en UML d'une classe d'objets flous à géométrie facultative

La transformation en UML d'une classe d'objets à géométrie facultative de type polygone flou introduit trois classes et un certain nombre de contraintes d'intégrité (cf. Figure 4). La classe *Ville* introduit toutes les villes qui seront stockées dans la base de données. La classe *PolygoneIntermediaire* ne présente que les villes floues qui satisfont les restrictions de l'utilisateur/fournisseur (objets dont on a l'information sur leur forme). La classe *Imperfection* définit les α-coupes représentant les villes floues (objets *PolygoneIntermédiaire*). La relation d'association entre la classe *Ville* et la classe *PolygoneIntermédiaire* présente la multiplicité (0, 1) : si une instance *Ville* satisfait la condition, elle est considérée comme un objet spatial flou et une forme de type polygone flou lui sera attribuée. Ainsi, un certain nombre d'α-coupes présentées par la classe *Imperfection* définit cette instance ; sinon, l'instance n'est pas considérée comme un objet spatial et n'a que les propriétés de la classe *Ville* (les attributs *nom*, *population*, *nbreLogement* et *nbreEtablissement*).

3.2. Patron de conception de géométrie alternative

Le modèle « *alternatif* » présente des géométries mutuellement exclusives. La géométrie d'une entité géographique peut être représentée par une dimension ou une autre. Plus précisément, une seule forme est numérisée pour chaque instance d'une classe d'objets géographiques. Dans ce cas, nous utilisons le pictogramme alternatif : c'est la concaténation des pictogrammes possibles ; il n'y a pas d'espace entre les

pictogrammes et leur ordre n'a aucun sens (Bedard, 1999). Le modèle « *alternatif* » présente un « OU EXCLUSIF » entre les pictogrammes.

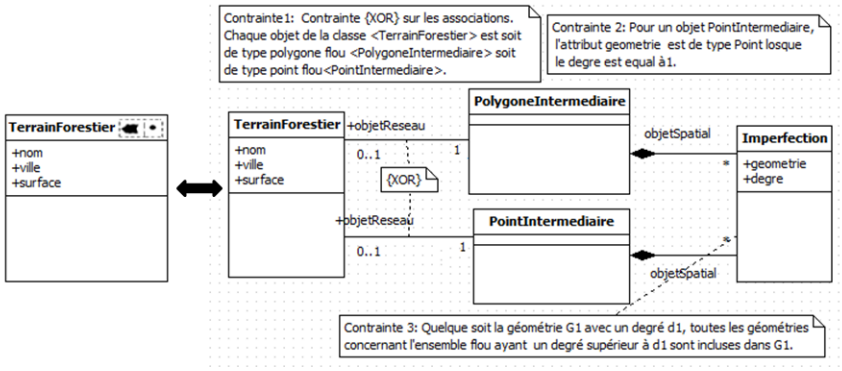


Figure 5. Dérivation en UML d'une classe d'objets flous à géométrie alternative

Prenant l'exemple d'un terrain forestier ; il peut être un parc urbain et il est numérisé ainsi par une géométrie de type point ou il peut être un parc régional et il est numérisé par géométrie de type polygone. En modélisation floue, la classe d'objets flous *TerrainForestier* est modélisée par deux pictogrammes : un polygone flou et un point flou (cf. Figure 5).

Dans la modélisation UML, nous définissons les classes *PolygoneIntermediaire* et *PointIntermediaire* respectivement pour les terrains forestiers de géométrie polygone flou et les terrains forestiers de géométrie point flou. Chaque instance *TerrainForestier* a deux formes possibles mais dont une seule forme est assignée. Cela veut dire que les deux associations s'excluent mutuellement. Cela est traduit par l'ajout de la contrainte $\{XOR\}$ entre les deux associations. La classe *Imperfection* présente les α-coupes de chaque objet flou (objet de type *PolygoneIntermédiaire* ou de type *PointIntermédiaire*).

3.3. Patron de conception de géométrie multiple

Le modèle « *multiple* » indique une situation où plusieurs formes sont numérisées pour chaque instance d'une classe d'objets géographiques (Bedard, 1999). Il peut être utilisé lorsqu'une seconde forme des instances d'objets peut être dérivée de leur première forme. Par exemple, une municipalité peut être modélisée par son centroïde et encore un polygone représentant ses frontières. Le centroïde peut être dérivé du polygone. Le modèle « *multiple* » est potentiellement utile lorsque les formes souhaitées ne peuvent être déduites l'une de l'autre, de sorte que la deuxième forme est dérivée à partir de la première forme de la même instance d'objet. Par exemple, une municipalité peut être modélisée par un centre-ville et un polygone : le centre-ville n'est pas dérivé du polygone (son centroïde).

Considérons l'exemple ci-après d'une classe d'objets flous *segment de route* dont la géométrie peut être modélisée avec une forme de type polygone si l'utilisateur veut montrer les frontières (à grande échelle) et une forme linéaire si l'utilisateur veut les représenter par leur lignes médianes (petite échelle).

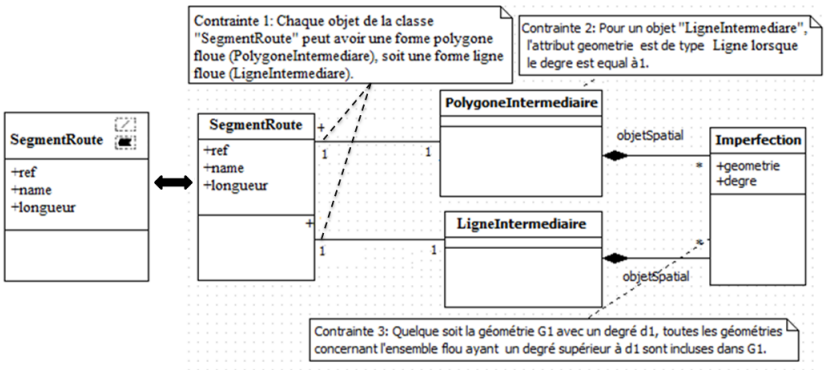


Figure 6. Dérivation en UML d'une classe d'objets flous à géométrie multiple

Pour chaque instance de segment de route floue, la géométrie est stockée dans la base de données à la fois par une forme de type polygone flou et une forme de type ligne floue.

La dérivation en UML de la classe *SegmentRoute* (cf. Figure 6) introduit des associations (un à un) entre la classe *SegmentRoute* et les classes { *LigneIntermediaire*, *PolygoneIntermediaire* }. Chaque objet *SegmentRoute* a une instance de type ligne floue et une instance de type polygone flou. Chaque objet *LigneIntermediaire* et *PolygoneIntermediaire* est composé de n α-coupes modélisées par la classe *Imperfection*. En ce qui concerne les contraintes d'intégrité, le modèle résultant définit toutes les contraintes discutées dans la section 2.4. En particulier, pour un segment de route représenté par une géométrie de type ligne floue (objet *LigneIntermediaire*), l'α-coupe a une géométrie de type ligne si le degré d'appartenance est égal à 1, sinon l'α-coupe a une géométrie de type polygone.

3.4. Patron de conception de géométrie agrégée

Le modèle « *agrégé* » présente des agrégations de géométries. La géométrie de la classe d'objets géographiques peut être une agrégation (1) de formes de même dimension (i.e. de formes similaires) ou (2) de formes de dimensions différentes (i.e. de formes hétérogènes).

Pour une agrégation de formes similaires (une agrégation simple selon (Bédard *et al.*, 2004)), au lieu d'utiliser des pictogrammes multiples de même forme, nous ajoutons la multiplicité (1, N) à côté du pictogramme choisi. Par exemple, un réseau routier flou est composé de plusieurs segments linéaires flous (cf. Figure 7). La classe

ReseauRoutier a un pictogramme 1-dimension (ligne floue) suivi d'une multiplicité (1, N).

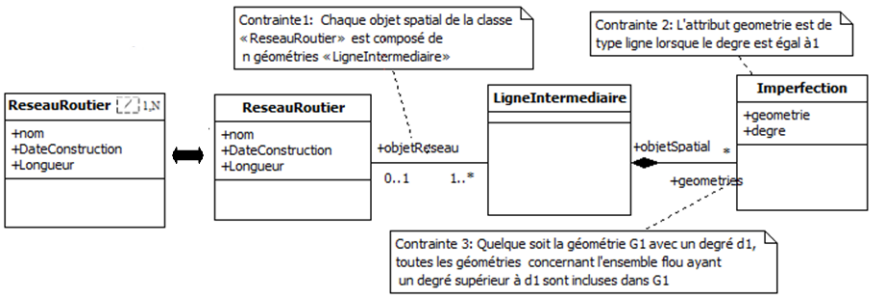


Figure 7. Dérivation en UML d'une classe d'objets flous à géométrie agrégée (cas d'agrégation monopictogrammique)

Pour une agrégation de formes hétérogènes (une agrégation complexe selon (Bédard *et al.*, 2004)), chaque instance est simultanément représentée par plusieurs formes. Par exemple, un réseau hydrographique flou est composé de rivières floues et de lacs flous. Le réseau hydrographique flou est modélisé par un pictogramme composé de polygone flou (pour modéliser les lacs) et de ligne floue (pour modéliser les rivières).

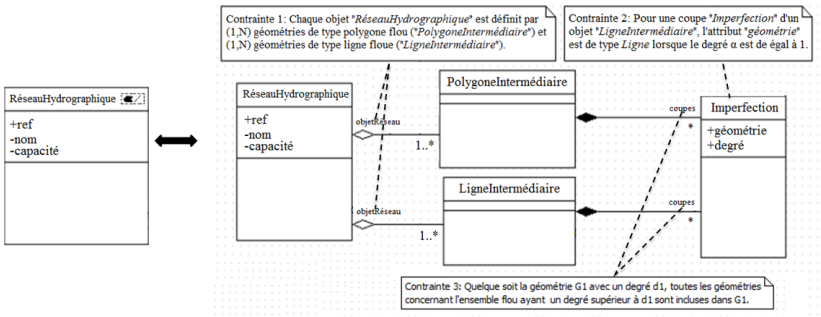


Figure 8. Dérivation en UML d'une classe d'objets flous à géométrie agrégée (cas d'agrégation multipictogrammique)

4. Exemple d'application

Nous avons mis en place un prototype de modélisation de bases de données géographiques floues baptisé F-Perceptory. Il est proposé sous l'outil logiciel Modelio dans lequel les utilisateurs peuvent créer, éditer et visualiser des diagrammes à travers un ensemble d'interfaces conceptuelles. Afin d'assurer la cohérence structurelle et sémantique de données floues dans la base de données, le prototype F-Perceptory est doté d'un processus de dérivation automatique de modèles, à partir du modèle conceptuel flou jusqu'à le script SQL. Basé sur un ensemble de règles de transformation, le

processus assure la transformation de toutes les spécifications de données structurelles et sémantiques des données géographiques floues et toutes les contraintes d'intégrité classiques sur les classes et les attributs.

4.1. Diagramme de classes floues

Pour créer un diagramme de classes floues, des menus contextuels et des boîtes de dialogue s'affichent à l'utilisateur lui permettant d'accéder à un ensemble d'actions possibles sur l'élément en cours (classe ou attribut), comme l'ajout de stéréotypes aux classes (à travers des pictogrammes renseignant sur le type de la géométrie ou la multiplicité), la modification et la suppression de stéréotypes, etc. Pour des contraintes d'implémentation, les pictogrammes se situent à droite du nom de la classe et les multiplicités s'ajoutent à gauche des pictogrammes.

La Figure 9 illustre un exemple de modélisation des réseaux hydrographiques à travers les villes. Les villes sont divisées en certaines régions hydrographiques dont chaque région contient un nombre de bassins versants. Chaque bassin versant peut avoir un réseau hydrographique. Le réseau hydrographique peut-être composé d'un certain nombre de cours d'eau (e.g. rivières, lignes de crête) et de zones exutoires (lacs).

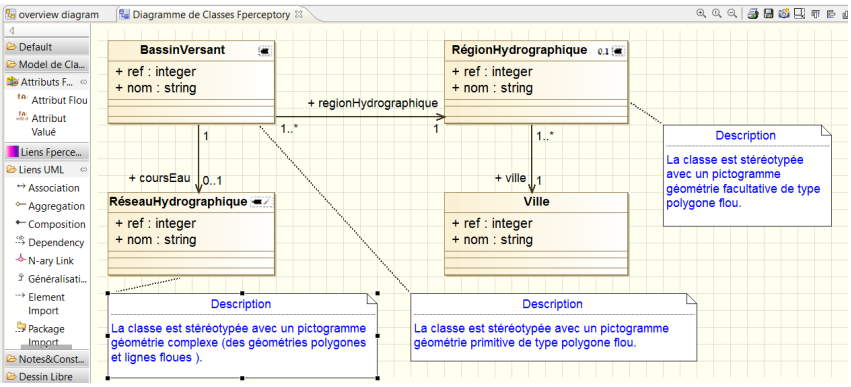


Figure 9. Diagramme de classes floues avec le prototype F-Perceptory

Le diagramme de classes floues F-Perceptory montre les régions hydrographiques ayant une géométrie facultative de type polygone flou (classe stéréotypée avec le pictogramme polygone flou et la multiplicité (0,1)), les bassins versants comme des espaces de type polygone flou (classe stéréotypée avec le pictogramme polygone flou), les réseaux hydrographiques comme des objets flous complexes (la classe est stéréotypée avec un pictogramme flou composé de polygone et de ligne) et enfin, les villes qui sont présentées en tant qu'entités non floues. Le diagramme dispose de deux classes avec des géométries composites floues, une classe avec une géométrie simple floue et une classe UML non stéréotypée.

4.2. Dérivation du diagramme de classes UML

Pour pouvoir gérer les données imprécises dans une base de données spatiales (objet-relationnelles), le modèle flou est transformé en un modèle UML (cf. Figure 10). En appliquant les règles de transformation discutées dans la section 2.4, cela implique la génération automatique de nouvelles classes et relations (en magenta) et d'un ensemble de contraintes d'intégrité exprimées dans le langage de contraintes orienté-objet OCL (*Object Constraint Language*)(en rouge). Notre objectif étant la transformation en UML standard, nous n'avons ni travailler avec l'OCL spatial, ni avec l'OCL spatial flou (Bejaoui *et al.*, 2010).

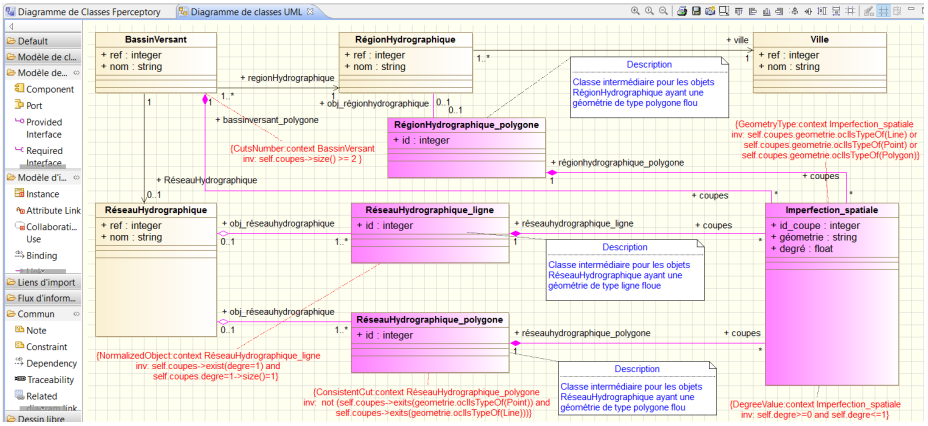


Figure 10. Diagramme de classes UML résultat de la transformation du diagramme de classes floues

Le diagramme UML ainsi généré comporte les éléments suivants :

– La classe *RéseauHydrographique* est une classe à géométrie complexe floue. En UML, elle agrège une classe intermédiaire à géométrie polygone flou (*RéseauHydrographique_polygone*) pour représenter les zones exutoires et une classe intermédiaire à géométrie ligne floue (*RéseauHydrographique_ligne*) pour représenter les cours d'eau.

– Les classes *RéseauHydrographique_polygone* et *RéseauHydrographique_ligne* sont floues. Elles ont chacune une relation de composition avec la classe *Imperfection_spatiale*. Cette relation permet de représenter les α-coupes de chaque objet flou définit dans les classes *RéseauHydrographique_polygone* et *RéseauHydrographique_ligne*.

– La classe *BassinVersant* est un polygone flou. Par conséquent, chaque objet bassin versant est composé de n α-coupes de type polygone ; elles sont représentées par la classe *Imperfection_spatiale*.

– La classe *RégionHydrographique* est un polygone flou facultatif. La classe *RégionHydrographique_polygone* présente les régions hydrographiques auxquelles est affectée une forme de type polygone flou.

– Les contraintes d'intégrité liées aux classes floues sont intégrées dans la vue "*notes and constraints*" de l'AGL. Certaines d'entre elles sont affichées dans le diagramme. Nous donnons un exemple illustré par la Figure 11 qui vérifie la contrainte de normalisation pour les objets cours d'eau représentés par la classe *RéseauHydrographique_ligne*. Chaque objet *RéseauHydrographique_ligne* doit avoir une seule α-coupe dont le degré α est égal à 1.

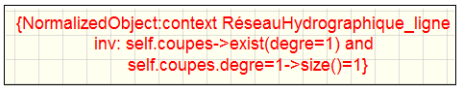


Figure 11. Contrainte OCL de la normalisation des instances de la classe RéseauHydrographique_ligne

À ce stade, il convient de noter que Modelio n'est pas basé sur un modèle de données spatiales, donc il ne gère pas les types de données spatiales comme *Geometry*, *Point*, *Polygon* et *Line*. L'attribut géométrique de la classe *Imperfection_spatiale* est, alors, mis par défaut avec le type de données chaîne de caractère (*String*). Nous proposons de le transformer en un type géométrique (*geography*) lors de la dérivation du schéma logique en script SQL pour le SGBD PostgreSQL/PostGIS.

4.3. Dérivation du schéma relationnel

Le prototype donne également la possibilité de générer le modèle relationnel de la base de données. Modelio fournit un module appelé « *PersistantProfile* » qui permet de dériver un modèle UML vers un schéma relationnel. A partir du schéma relationnel généré, nous pouvons ajouter d'autres contraintes classiques, comme par exemple, la contrainte d'unicité (*PrimaryKey*) pour chaque classe.

4.4. Dérivation du script SQL

Le prototype permet de dériver, à partir du schéma relationnel, le modèle de base de données physique sous forme d'un script SQL. C'est un script écrit en PL/pgSQL pour le SGBD spatial PostgreSQL/PostGIS. Cette dérivation permet de traduire automatiquement le type de données « *String* » de l'attribut *géometrie* de la classe *Imperfection_spatiale* en type de données « *Geography* ».

La dérivation automatique du schéma relationnel permet de générer un ensemble de tables et des contraintes d'intégrité qui peuvent être classiques (clés primaires, clés étrangères) ou des déclencheurs exprimés en langage procédural PL/SQL.

Nous donnons, ci-après, le résultat de la transformation de la contrainte de normalisation de chaque objet flou cours d'eau, appliquée sur la classe *RéseauHydrographique_ligne* (cf. Figure 12). Il s'agit d'un déclencheur qui vérifie que l'ensemble des α-coupes présentant chaque objet *RéseauHydrographique_ligne* contient une seule α-

coupe restrictive dont le degré d'appartenance = 1 et d'autres α-coupes ayant le degré d'appartenance $\neq 1$, sinon, le déclencheur renvoie une exception.

```
--Liste des triggers concernant la classe réseauhydrographique_ligne: nombre=4
--Trigger pour la contrainte NormalizedObject:
CREATE or replace FUNCTION verify_normalization()
RETURNS trigger AS $verify_normalization$
DECLARE
coupe Imperfection_spatiale%rowtype;
BEGIN
   SELECT into coupe
   FROM Imperfection_spatiale
   where (
     Imperfection_spatiale.réseauhydrographique_ligne = new.réseauhydrographique_ligne
     AND Imperfection_spatiale.degré = 1
  );
  IF NOT FOUND THEN
      RAISE EXCEPTION 'L''objet % n''est pas normalisé', new.réseauhydrographique_ligne;
   END IF;
RETURN NEW;
END;
$verify_normalization$ LANGUAGE plpgsql;
CREATE Trigger trigger_NormalizedObject AFTER INSERT ON Imperfection_spatiale
                FOR EACH ROW  EXECUTE PROCEDURE verify_normalization();
...
```

Figure 12. Déclencheur PL/SQL pour la vérification de la normalisation des instances de la classe RéseauHydrographique_ligne

5. Conclusion

Nous avons présenté une extension de F-Perceptory pour la modélisation des géométries composites. La version étendue est mise en œuvre à travers un prototype afin de supporter les utilisateurs dans la modélisation conceptuelle de données floues. Un processus de dérivation automatique de modèles est également implémenté afin de générer le modèle de la base de données spatiales pour le système PostgreSQL/PostGIS.

Nous avons accordé, dans ce papier, une attention particulière à la spatialité floue. Nous envisageons de compléter notre démarche par l'ajout de la modélisation de la temporalité floue et par une étude pour la considération dans les modèles des relations topologiques entre objets flous.

Bibliographie

Bedard Y. (1999). Visual modelling of spatial databases : towards spatial PVL and UML. *Geomatica*, vol. 53, n° 2, p. 169–186.

Bédard Y., Larrivée S., Proulx M.-J., Nadeau M. (2004). Modeling geospatial databases with plug-ins for visual languages: A pragmatic approach and the impacts of 16 years of research and experimentations on Perceptory. In *Proceedings of international conference on conceptual modeling*, p. 17–30.

Bedard Y., Proulx M.-J., Larrivée S., Bernier E. (2002). Modeling multiple representations into spatial data warehouses : a UML-based approach. In *Symposium sur la Théorie, les Traitements et les Applications des Données Géospatiales.*

Bejaoui L., Pinet F., Schneider M., Bédard Y. (2010). Ocl for formal modelling of topological constraints involving regions with broad boundaries. *GeoInformatica*, vol. 14, n° 3, p. 353–378.

Chen G., Kerre E. (1998). Extending ER/EER concepts towards fuzzy conceptual data modeling. In *Proceedings of IEEE International Conference on Fuzzy Systems*, vol. 2, p. 1320–1325.

De Ruffray S. (2007). *L'imprécision et l'incertitude en géographie. l'apport de la logique floue aux problématiques de régionalisation*. Mémoire d'habilitation à diriger des recherches. Université Paris Diderot-Paris 7.

Desjardin É., Runz C. de, Pargny D., Nocent O. (2012). Modélisation d'un SIG archéologique et développement d'outils d'analyse prenant en compte l'imperfection de l'information. *Revue Internationale de Géomatique/International Journal of Geomatics and Spatial Analysis*, vol. 22, n° 3, p. 367–387.

Ma Z. (2005). Fuzzy information modeling with the UML. In *Advances in fuzzy object oriented databases : Modeling and applications*, p. 153-176. Elsevier.

Ma Z., Zhang W.-J., Ma W., Chen G. (2001). Conceptual design of fuzzy object-oriented databases using extended entity-relationship model. *International Journal of Intelligent Systems*, vol. 16, n° 6, p. 697–711.

Miralles A. (2006). *Ingénierie des modèles pour les applications environnementales*. Thèse de doctorat non publiée, Université Montpellier, France.

Pantazis D., Donnay J. (1996). *Conception des SIG: Méthode et formalisme*. Paris : Hermès et Lavoisier.

Pantazis D., Donnay J.-P. (1998). Objets géographiques à limites indéterminées. modélisation et intégration dans un modèle conceptuel de données. *Revue Internationale de Géomatique*, vol. 7, n° 2, p. 159–186.

Parent C., Spaccapietra S., Zimányi E., Donini P., Plazanet C., *et al.* (1997). MADS : un modèle conceptuel pour des applications spatio-temporelles. *Revue Internationale de Géomatique*, vol. 7, n° 3, p. 317–352.

Shu H., Spaccapietra S., Sedas D. Q. (2003). Uncertainty of Geographic Information and its Support in MADS. In *Proceedings of International Symposium on Spatial Data Quality (ISSDQ)*, vol. 2.

Sicilia M.-A., Garcia-Barriocanal E. (2006). Extending object database interfaces with fuzziness through aspect-oriented design. *ACM Special Interest Group on Management of Data Record (SIGMOD Record)*, vol. 35, n° 2, p. 4–9.

Yazici A., Akkaya K. (2000). Conceptual modeling of geographic information system applications. In *Recent Issues on Fuzzy Databases*, p. 129–151. Springer.

Zadeh L. A. (1965). Fuzzy sets. *Information and Control*, vol. 8, n° 3, p. 338–353.

Zayrit K., Desjardin É. (2012). Formalisation des imperfections dans les données dans d'un système d'information agro-environnemental : Observox. In *Spatial analysis and geomatics.* Liège, Belgique.

Zoghlami A. (2016). F-Perceptory : an approach for handling fuzziness of spatiotemporal data in geographical databases. *International Journal of Spatial, Temporal and Multimedia Information Systems (IJSTMIS)*, vol. 1, n° 1, p. 30–62.

Zvieli A., Chen P. P. (1986). Entity-relationship modeling and fuzzy databases. In *Proceedings of IEEE International Conference on Data Engineering*, vol. 2, p. 320–327.

Analyse de l'information dans les réseaux sociaux

La qualité de l'information dans les réseaux sociaux en ligne: une approche non supervisée et rapide de détection de spam

Mahdi Washha , Manel Mezghani , Florence Sèdes

Institut de Recherche en Informatique de Toulouse (IRIT), Université de Toulouse, CNRS, INPT, UPS, UT1, UT2J, 31062 TOULOUSE Cedex 9, France
mahdi.washha,manel.mezghanni,florence.sedes@irit.fr

RÉSUMÉ. *Les réseaux sociaux en ligne fournissent des données utiles pour une vaste gamme d'applications. Cependant, les interfaces faciles à utiliser et les faibles limites de sécurité à la publication génèrent divers problèmes liés à la qualité de "l'information", i. e. des contenus générés par l'utilisateur via ces réseaux. L'existence d'utilisateurs mal intentionnés, nommés spammeurs sociaux, entre dans ce cadre. Les principales limitations des approches de détection de spam sont qu'elles reposent sur des approches d'apprentissage supervisé qui exigent des ensembles de données vérité terrain. Par ailleurs, les approches de détection basées sur les comptes des utilisateurs ne sont pas utilisables pour le traitement de grandes collections de publications, car elles nécessitent des mois pour traiter ces collections. Dans cet article, nous présentons donc une approche d'apprentissage non supervisé dédiée à la détection de comptes spam dans de grandes collections de thématiques "tendances". Notre approche a été testée sur Twitter et a prouvé son efficacité ainsi que sa rapidité par rapport aux approches d'apprentissage supervisé.*

ABSTRACT. *Online social networks provide data valuable for a tremendous range of applications. However, the easy-to-use interfaces and low limits of the publication generate various information quality problems, such the user-generated content in such networks. The existence of ill-intentioned users, so-called social spammers, belongs to this environments. The major limitations of the methods detecting spams are the use of supervised learning approaches that requiring ground truth data-sets. Moreover, the account-based detection methods are not practical for processing "crawled" large collections of social posts, requiring months to process such collections. Hence, in this paper, we introduce a design of an unsupervised learning approach dedicated for detecting spam accounts existing in large collections of "trending" topics. Our experimental evaluation on Twitter demonstrates the efficiency of our approach as well as its speed comparing to supervised approaches.*

MOTS-CLÉS : *Twitter, Réseau Social, Spam*

KEYWORDS: *Twitter, Social Network, Spam*

1. Introduction

Les réseaux sociaux en ligne (OSN) sont devenus un moyen de communication puissant dans lequel les utilisateurs ont la possibilité de partager des liens, de discuter et de s'inter-connecter. Les interfaces faciles à utiliser et les faibles limites de sécurité à la publication ne contribuent pas à maintenir un niveau constant de qualité de l'information (QI). Ces caractéristiques ont rendu les OSN vulnérables à diverses attaques par un certain type d'utilisateurs mal intentionnés, appelés spammeurs sociaux. Les spammeurs sociaux affichent un contenu illicite ou non pertinent par rapport à un contexte donné ou une thématique particulière. Plus généralement, les spammeurs sociaux ont à disposition un large éventail tde echniques pour publier des contenus spam, résumés dans (Benevenuto *et al.*, 2010): (i) diffusion de publicités pour générer des ventes et des profits illégaux; (ii) diffusion de matériel pornographique; (iii) publication de virus et de malwares; (iv) création de sites Web de phishing pour révéler des informations sensibles, ...

Les impacts négatifs du spam social ne pouvant être ignorés: traiter ce problème contribue à améliorer la qualité de l'information en détectant et en filtrant les contenus "spam" existants dans les collections de données téléchargées ou dans les OSN. La présence du spam dans les OSN est à l'origine de nuisances majeures telles que: (i) polluer les résultats de recherche; (ii) dégrader l'exactitude des statistiques obtenues à travers des outils d'extraction d'information; (iii) consommer des ressources de stockage; (iv) violer la vie privée des utilisateurs, ... Les mécanismes anti-spam s'avèrent insuffisants pour mettre fin au problème de spam, ce qui suscite de réelles inquiétudes quant à la qualité des collections de données "aspirées". La solution proposée dans ce papier peut être intégrée à des recherches portant sur un large éventail de problèmes de OSN afin d'améliorer leurs résultats, comme par exemple les travaux de (Abascal-Mena *et al.*, 2015 ; Canut *et al.*, 2015) sur le profilage social et la détection des communautés socio-sémantiques, dans lesquesl des collections de données à grande échelle des thématiques tendances de Twitter servent de base aux expérimentations.
Dans la bataille de la détection du spam social sur Twitter, plusieurs approches ont été proposées pour détecter les campagnes de spam et les comptes spam individuels, mais peu d'efforts ont été dédiés aux tweets spam individuels. Ces approches fonctionnent en utilisant le concept d'extraction de caractéristiques combiné avec une approche d'apprentissage supervisé pour construire un modèle prédictif basé sur un jeu de données annoté (vérité terrain). Cependant, s'appuyer sur l'approche d'apprentissage supervisé dans la conception de modèles prédictifs devient moins efficace pour détecter les utilisateurs spammeurs ou du contenu spam en raison de l'évolution rapide du comportement de ces spammeurs sociaux. Ceux-ci adoptent des modèles de contenu dynamique dans les OSNs ainsi qu'un changement de leurs stratégies de spam pour se faire passer pour des utilisateurs "normaux". Les approches anti-spam statiques présentent donc un retard considérable en ne considérant pas l'évolution rapide des comportements des spammeurs sociaux.
Au-delà de cette dynamique évidente, les approches de détection actuelles basées sur les campagnes et sur les comptes individuels ne conviennent pas aux ensembles de données à grande échelle de thématiques de Twitter, nécessitant des mois pour traiter

ce volume de données. Comme les tweets se composent de méta-données simples (par exemple nom d'utilisateur, date de création), ces approches ont été conçues en fonction de caractéristiques avancées (par exemple, les tweets de l'utilisateur au fil du temps) nécessitant des informations supplémentaires des serveurs de Twitter. Ce dernier fournit des fonctions pour récupérer ces informations supplémentaires (par exemple, les abonnés de l'utilisateur - *followers* et les *followees*) sur un objet donné à l'aide des API REST [1]. Cependant, Twitter limite le nombre d'appels autorisés via l'API à une fenêtre de temps définie (par exemple, 40 appels en 15 minutes). La récupération d'informations, y compris les méta-données des *followers* et des *followees*, pour un demi-million d'utilisateurs qui ont posté un million de tweets, peut dès lors prendre environ trois mois.

Dans cet article, nous proposons une approche non supervisée pour filtrer les comptes spam dans des ensembles de données à grande échelle de thématiques tendances dans Twitter. Plus précisément, nous n'exploitons que les méta-données disponibles dans une thématique donnée, sans récupérer aucun type d'information des serveurs de Twitter.

Le reste de l'article est organisé comme suit. La section 2 présente le mécanisme anti-spam de Twitter ainsi que les approches de détection de spam social proposées dans la littérature. La section 3 introduit les notations, la formalisation des problèmes et la conception de notre approche. La section 4 détaille l'ensemble de données utilisées pour l'expérimenter et la valider. La configuration expérimentale et une série d'expérimentations évaluant l'approche proposée sont décrites dans la section 5. Enfin, la section 6 conclut le travail.

2. Contexte et Etat de l'art

Dans cette section, nous présentons le mécanisme utilisé dans Twitter pour combattre le spam ainsi que les approches de détection de spam.

Mécanisme Anti-Spam de Twitter. Twitter combat les spammeurs sociaux en permettant aux utilisateurs de signaler des comptes spam simplement en cliquant sur l'option "Signaler: ils publient du spam" disponible sur la page du compte. Lorsqu'un utilisateur signale un compte particulier, les administrateurs de Twitter examinent manuellement le compte rendu pour prendre la décision de suspension. Cependant, l'adoption d'une telle approche pour lutter contre les spammeurs nécessite des efforts considérables des utilisateurs et des administrateurs. Malheureusement, la probabilité de détecter et de suspendre un compte quelques jours après sa création est inférieure à 1%. Par conséquent, ces lacunes ont motivé les chercheurs à proposer des approches plus puissantes pour les applications qui utilisent Twitter comme source d'information à savoir les approches d'apprentissage automatique (*Machine learning approaches*) comme étant une approche entièrement automatisée que nous détaillons dans ce papier.

Approche d'apprentissage automatique. Ces approches sont construits en employant trois niveaux de détection:

1. https://dev.twitter.com/rest/public

Niveau tweet: À ce niveau, les tweets individuels sont vérifiés afin d'éliminer d'éventuels contenus indésirables. Benevenuto (Benevenuto *et al.*, 2010) a extrait un ensemble de caractéristiques statistiques simples du tweet telles que le nombre de mots, le nombre de hashtags et le nombre de caractères. Ensuite, un classifieur binaire est construit sur un petit ensemble de données annotées. Martinez-Romo et Araujo (Martinez-Romo, Araujo, 2013) ont détecté des tweets spam dans les thématiques tendances à travers l'utilisation de modèles de langage pour extraire plus de caractéristiques telles que la divergence de distribution de probabilité entre un tweet donné et d'autres tweets. Le problème majeur à ce niveau de détection provient du manque d'information qui peut être extraite du tweet lui-même. En outre, la construction de modèles de langues à l'aide de tweets dans les thématiques tendances échoue définitivement quand il y a d'énormes attaques de spam.

Niveau compte: Les approches conçues dans (Wang, 2010 ; Benevenuto *et al.*, 2010 ; Stringhini *et al.*, 2010 ; McCord, Chuah, 2011 ; Cao, Caverlee, 2015) construisent d'abord des vecteurs en extrayant des caractéristiques "à la main" telles que le nombre de *followers*, et l'intermédiation de noeuds. Ensuite, des algorithmes d'apprentissage automatique supervisés sont appliqués pour construire un modèle de classification sur un ensemble de données annotées. Malgré un taux de détection élevé en exploitant ces caractéristiques, les extraire est chronophage en raison du temps nécessaire au recueil des informations du serveur de Twitter via l'utilisation de l'API REST. En effet, ces API sont limitées à un certain nombre prédéfini d'appels, ce qui rend l'extraction de la plupart des caractéristiques impossible, en particulier dans le traitement de données à grande échelle.

Niveau campagne: Chu et al. (Chu, Widjaja, Wang, 2012) ont proposé une approche de détection de campagne de spam à travers le regroupement des comptes spam selon les URL disponibles dans les tweets. Un vecteur est ensuite représenté, via des caractéristiques similaires aux approches de détection au niveau du compte. Dans (Chu, Gianvecchio *et al.*, 2012), un modèle de classification a été conçu pour capturer les différences entre campagne, humain et *cyborg*. Malheureusement, ce niveau de détection présente des inconvénients similaires à ceux mentionnés pour le niveau "compte", ce qui rend ces solutions non évolutives pour de grandes collections d'utilisateurs ou de tweets.

Nous présentons dans la section suivante notre approche afin de surmonter les problématiques évoquées dans l'apprentissage automatique au niveau *tweet* .

3. Modèle Prédictif

Dans cette section, nous présentons les premières définitions et notations utilisées dans la modélisation de notre solution. Ensuite, nous présentons la conception de notre approche de détection des comptes spam.

3.1. Notations, définitions et formalisation du problème

Le concept de *Topic* [2] ou thématique peut être défini comme une représentation de structures sémantiques cachées dans une collection de textes (par exemple: des documents textuels, des tweets). *Trending Topic* ou thématique tendance est un mot ou une expression (par exemple #TopChef) qui est mentionné à un taux plus élevé que d'autres. Les thématiques tendances sont automatiquement identifiées par un algorithme qui identifie des thématiques qui sont plus massivement diffusées par les utilisateurs que d'autres.

Comme plusieurs utilisateurs peuvent tweeter sur le même sujet, nous modélisons une thématique tendance particulière T comme un ensemble fini d'utilisateurs distincts, définie comme $Topic(U_T) = \{u_1, u_2, ...\}$, où l'élément utilisateur (ou compte) $u_\bullet$ est en outre défini par un 4-tuple d'attributs, $u_\bullet = (UN, SN, UA, TS)$. Chaque attribut dans l'élément utilisateur est défini comme suit:

Nom d'utilisateur ou *User Name* (UN): Nous modélisons cet attribut comme un ensemble de caractères ordonnés, définis comme $UN = \{(1, a_1), ..., (i, a_i)\}$, où $i \in \mathbb{Z}_{\geq 0}$, est la position dans la chaîne de nom d'utilisateur et $a_\bullet \in \{Printable\,Characters\}$ [3], est le caractère.

Nom dans l'écran ou *Screen Name* (SN): De la même façon que l'attribut *username*, nous modélisons ce champ comme un ensemble ordonné de caractères, défini comme $SN = \{(1, a_1), ..., (i, a_i)\}$ où $i \in \mathbb{Z}_{\geq 0}$ est la position à l'intérieur de la chaîne de noms d'écran, $a_\bullet \in \{Alphanumeric\,Characters\} \cup \{\_\}$, est le caractère.

Age utilisateur ou *User Age* (UA): Formellement nous calculons l'âge dans l'unité de temps *days* en soustrayant l'heure courante de la date de création du compte, défini comme $UA = \frac{Time_{now} - Time_{creation}}{864 * 10^5}$, où $Time_{now}, Time_{creation} \in \mathbb{Z}_{\,ge0}$ sont le nombre de millisecondes calculé depuis le 1er janvier 1970, 00:00:00 GMT.

Ensemble de tweet ou *Tweets Set* (TS): Chaque utilisateur $u_\bullet \in Topic(U_T)$ peut publier plus d'un tweet dans la rubrique T, où chaque tweet peut décrire les pensées ou intérêts de l'utilisateur ou intérêts. Ainsi, nous modélisons ce champ comme un ensemble fini de tweets, définis comme $TS = \{TW_1, ..., TW_n\}$, où n représente le nombre de tweets publiés par l'utilisateur $u_\bullet$. De plus, nous modélisons chaque tweet en double-tuple d'attributs, $TW_\bullet = (Text, Time)$, où $Text = \{(1, w_1), (2, w_2), ...\}$ est un ensemble fini de mots de chaîne, et $Time \in\ mathbbZ_{\geq 0}$ est la date de publication du tweet dans l'unité de temps *minutes* calculée depuis le 1er janvier 1970, 00:00:00 GMT.

Formalisation du problème. Etant donné un ensemble d'utilisateurs U_T, tels que chaque utilisateur a posté un ou plusieurs tweets dans une thématique tendance T, notre problème principal est de prédire le type (spam ou non-spam) de chaque utilisateur dans l'ensemble donné U_T, sans nécessiter aucune connaissance préalable telle que la rela-

2. https://support.twitter.com/articles/101125
3. http://web.itu.edu.tr/sgunduz/courses/mikroisl/ascii.html

tion entre les utilisateurs (par exemple les *followers* et les *followees* des utilisateurs). Plus formellement, nous cherchons à concevoir une fonction y qui traite et gère l'ensemble donné d'utilisateurs U_T, pour prédire l'étiquette de chaque classe d'utilisateurs, définie comme $y : u_\bullet \rightarrow \{spam, non - spam\}$, $u_\bullet \in U_T$.

3.2. Modèle à quatre étapes

Nous concevons une approche à quatre étapes illustrée par la figure 1 et détaillée ci-dessous.

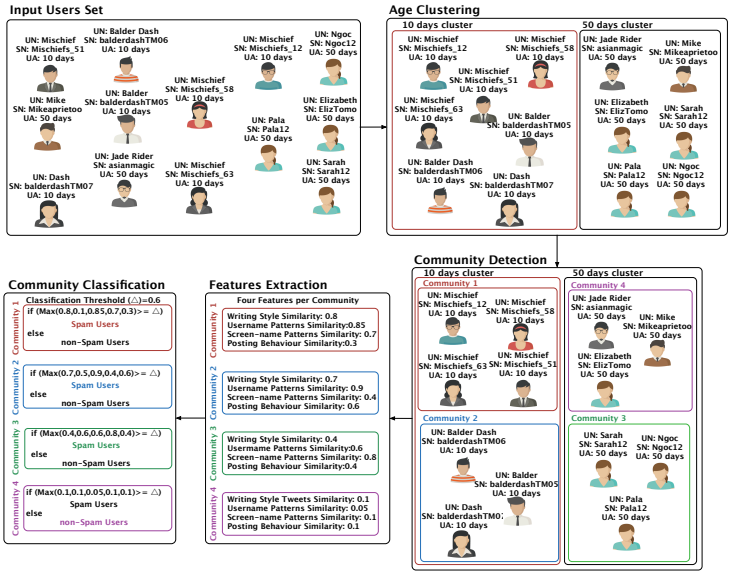


Figure 1. Un exemple détaillant les quatre étapes du modèle proposé, depuis l'ensemble d'utilisateurs qui a posté des tweets dans une thématique jusqu'à leur classification

3.2.1. Regroupement en fonction de l'âge des utilisateurs

Les spammeurs sociaux ont la capacité de créer des centaines et des milliers de comptes Twitter dans un court laps de temps ne dépassant pas quelques jours, pour lancer leurs campagnes de spam. Dans ce cas, la date de création des comptes peut contribuer à regrouper et à isoler les comptes spam qui pourraient avoir une corrélation entre eux. Par conséquent, nous regroupons les utilisateurs en fonction de l'attribut âge de l'utilisateur (UA). D'une manière formelle, soit $C_{age} = \{u|U \in U_T, u.UA = age\}$ un groupe ou *cluster* contenant les utilisateurs ayant un âge égal à $age \in Ages$ où $Ages = \{u.UA|U \in U_T\}$ est un ensemble d'âges d'utilisateurs. De toute évidence, le nombre de groupes d'âges est exactement égal à la taille de $Ages$ (c'est-à-dire $|Ages|$).

3.2.2. Détection de communauté

Nous exploitons l'utilisation de l'approche de factorisation matricielle non négative ou *non-negative matrix factorization* (NMF) (Yang, Leskovec, 2013), comme une

approche non supervisée, pour déduire la structure des communautés en raison de sa remarquable performance dans le regroupement. Cette approche partitionne une matrice d'information en matrices de facteurs cachés pour un groupe d'âge C_{age}, d'utilisateurs, défini mathématiquement comme un problème de minimisation d'optimisation:

$$\min_{H \geq 0} ||\mathbf{X} - \mathbf{H}\mathbf{H}^T||_F^2 \quad (1)$$

Où $|| \bullet ||_F$ est la norme de Frobenius de la matrice considérée, $\mathbf{X} \in R^{|C_{age}| \times |C_{age}|}$ est une matrice d'information représentant la force des connexions sociales entre les utilisateurs, $\mathbf{H} \in R^{|C_{age}| \times K}$ est la matrice des facteurs de la structure communautaire de K. L'entrée $X(i,j)$ (i et j représentent les indexes dans la matrice) reflète la force du lien social entre l'utilisateur $u_i \in C_{age}$ et l'utilisateur $u_j \in C_{age}$. L'entrée $H(i,j)$ (i et j représentent les indexes dans la matrice) dans la matrice de facteur caché peut être interprétée comme le degré de confiance de l'utilisateur $u_i \in C_{age}$ appartenant à la communauté j^{th}. Il est important de mentionner que chaque utilisateur appartient à une seule communauté uniquement.

Évidemment, inférer la matrice cachée **H** requiert une définition formelle de la matrice d'information **X**. Par exemple, **X** peut être une matrice d'adjacence représentant les liens sociaux ou les liens entre les utilisateurs du groupe d'âge donné C_{age}. Cependant, l'obtention de la matrice d'adjacence dans notre cas n'est pas possible car les informations disponibles sur les utilisateurs sont limitées à des méta-données simples décrivant des comptes sans fournir d'informations sur les *followers* et les *followees*. Par conséquent, dans cet article, nous mettons à profit l'information disponible et accessible pour estimer les connexions sociales entre les utilisateurs en proposant deux définitions de la matrice d'information **X** notée $\mathbf{X}^{SN}$ et $\mathbf{X}^{UN}$, dont chacun est formellement défini comme suit:

Name ($\mathbf{X}^{SN}$): Comme le champ SN doit être unique, les spammeurs ont tendance à adopter un modèle fixe particulier lors de la création de plusieurs comptes pour agir comme des campagnes de spam. Intuitivement, le chevauchement ou l'appariement élevé dans le SN parmi les utilisateurs augmente la probabilité pour les utilisateurs d'appartenir à la même communauté. Par conséquent, nous définissons la matrice d'information $\mathbf{X}^{SN}$ pour mesurer le degré d'appariement dans l'attribut SN. Plus précisément, étant donné deux utilisateurs $u_i, u_j \in C_{age}$, le degré d'appariement pour une entrée particulière dans la matrice $\mathbf{X}^{SN}$ est défini comme suit:

$$\mathbf{X}^{SN}(i,j) = \frac{max\{|m| : m \in N - gram(u_i.SN) \cap N - gram(u_j.SN), N \in Max\}}{min(|u_i.SN|, |u_j.SN|)}$$

Où $|\bullet|$ est la cardinalité de l'ensemble considéré, $Max = \{1, ..., min(|u_i.SN|, |u_j.SN|)\}$ est un ensemble composé d'entiers positifs représentant le nombre potentiel de caractères qui se chevauchent entre les noms donnés, $N - gram(\bullet)$ est une fonction qui renvoie un ensemble de séquences contiguës de caractères pour le nom donné (ensemble de caractères ordonnés) sur la base de la valeur de N. Pour une meilleure compréhension, le 3-gramme (ou le trigramme) d'un SN égale à "vote" est $\{\{(1,v),(2,o),(3,t)\}, \{(1,o),(2,t),(3,e)\}\}$. La définition ci-dessus peut détecter le

motif assorti partout où il apparaît dans l'attribut SN. Par exemple, «vote12» et «tovote» sont des SN pour deux utilisateurs différents de spam, le degré de correspondance selon l'équation 3.2.2 est autour de $(\frac{4}{6})66.6\%$, résultant de l'utilisation du motif «vote», quelle que soit la position du motif.

Similarité de nom d'utilisateur ($\mathbf{X}^{UN}$): Contrairement à l'attribut SN, les spammeurs peuvent copier l'attribut de nom d'utilisateur autant qu'ils le souhaitent. Ils exploitent des noms représentatifs (pas aléatoires) pour attirer les utilisateurs non-spam. Par conséquent, la correspondance totale ou partielle entre les utilisateurs dans cet attribut augmente les performances de détection de la communauté. Nous définissons la matrice d'information $\mathbf{X}^{UN}$ pour mesurer le degré de similarité entre les utilisateurs dans l'attribut de nom d'utilisateur. Formellement, étant donné deux utilisateurs $u_i, u_j \in C_{age}$, le degré de similarité est défini comme suit:

$$\mathbf{X}^{UN}(i,j) = \frac{max\{|m| : m \in N-gram(u_i.UN) \cap N-gram(u_j.UN), N \in Max\}}{min(|u_i.UN|, |u_j.UN|)} \tag{2}$$

Où $Max = \{1, ..., min(|u_i.UN|, |u_j.UN|)\}$.

Combiner les informations des matrices. Avec ces deux matrices d'information, l'approche NMF permet de les intégrer dans la même fonction objective. Ainsi, la nouvelle version de la fonction objectif est définie comme suit:

$$\min_{H \geq 0} ||\mathbf{X}^{SN} - \mathbf{H}\mathbf{H}^T||_F^2 + ||\mathbf{X}^{UN} - \mathbf{H}\mathbf{H}^T||_F^2 \tag{3}$$

De toute évidence, l'équation 3 infère la matrice de facteurs cachés H pour représenter la structure communautaire cohérente des utilisateurs.

Approche d'optimisation. La fonction objective n'est pas conjointement convexe et aucune solution de forme fermée existe. Par conséquent, nous proposons l'utilisation de la descente en gradient comme une autre approche d'optimisation (Yang, Leskovec, 2013). Comme nous avons une variable libre de matrice (**H**), l'approche de descente par gradient la met à jour itérativement jusqu'à ce que la variable converge. Nous exploitons deux conditions d'arrêt: (i) le nombre d'itérations, noté M; (ii) et la variation absolue de la matrice H dans deux itérations consécutives est inférieure à un seuil, c'est-à- dire $|(||H^{\tau}||_F - ||H^{\tau-1}||_F)| \leq \epsilon$.

3.2.3. Extraction des caractéristiques basées sur la communauté

Afin de classer chaque communauté, un ensemble de caractéristiques peut être extrait pour discriminer parmi les communautés spam ou non-spam. Aucune caractéristique n'est capable de discriminer efficacement entre les communautés de spam et non-spam. De plus, la conception des caractéristiques doit maintenir la condition que la récupération d'informations sur les utilisateurs de serveurs Twitter n'est pas autorisée du tout. Par conséquent, nous proposons une conception de quatre caractéristiques qui examinent la perspective collective des utilisateurs, en utilisant uniquement les informations disponibles sur les utilisateurs. Nos caractéristiques sont réparties entre les caractéristiques basées sur le tweet et celles basées sur l'utilisateur,

répertoriées comme: *username patterns similarity* (**UNPS**), *screen-name patterns similarity* (**SNPS**), *tweets writing style similarity* (**TsWSS**) et *tweets posting behavior similarity* (**TPBS**). Nous formalisons la j^{eme} communauté extraite comme sextuple d'attributs, $C_j = \{U, UNPS, SNPS, TsWSS, TPBS, L\}$ où U est un ensemble d'utilisateurs appartenant à la j^{eme} communauté qui peut être extraite de **H** matrix, et $L \in \{spam, non-spam\}$ est l'étiquette de la j^{eme} communauté.

UNPS et SNPS: Les spammeurs peuvent adopter un modèle pour créer leurs campagnes de spam. Ainsi, lorsqu'une communauté est biaisée par rapport à un modèle particulier utilisé dans la création de comptes, cette communauté a une forte probabilité d'être une campagne de spam et, par conséquent, tous les utilisateurs de cette communauté sont des utilisateurs spam (comptes). Nous modélisons ce comportement de spam en trouvant d'abord tous les modèles possibles utilisés pour nommer des comptes, extraits des attributs UN et SN. Ensuite, nous calculons la distribution de probabilité de motifs extraits de UN ou SN. Enfin, nous comparons la distribution de probabilité calculée d'un attribut avec la distribution uniforme des motifs extraits. En effet, nous émettons l'hypothèse que la distribution de probabilité des modèles de communautés non-spam doit être proche de la distribution uniforme.

De manière formelle, soit PT^{UN} et PT^{SN} deux ensembles de chaînes de caractères finis extraits des attributs UN et SN de la j^{eme} communauté, respectivement. De même, soit P_D^{UN} et P_D^{SN} les distributions de probabilité des patrons d'attributs UN et SN, respectivement. Pour la distribution uniforme, soit $P_{uniforme}$ une distribution uniforme correspondante des modèles d'attributs considérés. Par exemple, dans le cas d'une communauté donnée, notons $PT^{SN} = \{"mischief", "isch", "\_12", "\_14"\}$, $P_D^{SN} = \{("mischief",\ 0.7), ("\_15", 0.1), ("\_14", 0.1), ("\_12", 0.1)\}$ un ensemble de motifs de SN avec sa distribution de probabilité uniforme. La distribution de probabilité de ces modèles est $\{("mischief", 0.25), ("\_15", 0.25), ("\_14", 0.25), ("\_12", 0.25)\}$.

Pour extraire des chaînes de caractères à partir d'un UN ou d'un SN, nous appliquons l'approche des caractères N-gram sur le UN ou le SN dans la communauté inférée. Comme les spammeurs peuvent définir des modèles variant en longueur et en position, pour prendre tous ou la plupart des modèles possibles, nous utilisons différentes valeurs de N allant de trois à la longueur du nom. Nous évitons $N \in \{1, 2\}$ car il est inutile d'avoir un ou deux caractères de modèle. Formalement, pour un nom donné sous la forme de $Name = \{(1, a_1), (2, a_2), ...\}$, dont les profils potentiels sont extraits par:

$$Patterns(Name) = \bigcup_{N \in Max} N - gram(Name) \tag{4}$$

Où $Max = \{3, ..., \mid Name \mid\}$ est un ensemble fini de valeurs possibles de N.

Avec la définition introduite, les ensembles de modèles de chaînes de la j^{eme} communauté sont donnés comme suit:

$$PT^{UN} = \bigcup_{u \in c_j \cdot U} Patterns(u \cdot UN), \quad PT^{SN} = \bigcup_{u \in c_j \cdot U} Patterns(u \cdot SN) \tag{5}$$

Nous quantifions la similarité entre les distributions en effectuant une corrélation croisée entre la distribution de probabilité des modèles associés à une communauté et la distribution uniforme correspondante aux modèles considérés. Formalement, nous calculons la similarité de distribution de probabilité pour les attributs UN et SN à l'aide des formules suivantes:

$$UNPS(C_j) = 1 - \frac{Area(P_D^{UN} \star P_{uniform})}{Area(P_{uniform} \star P_{uniform})} \tag{6}$$

$$SNPS(C_j) = 1 - \frac{Area(P_D^{SN} \star P_{uniform})}{Area(P_{uniform} \star P_{uniform})} \tag{7}$$

Où $Area\,(\bullet)$ est une fonction qui calcule la zone sous la nouvelle distribution résultante par l'opération de corrélation. L'idée-clé d'effectuer l'auto-corrélation entre la distribution de probabilité uniforme et elle-même, est de normaliser la valeur de la zone qui provient de l'opération de corrélation croisée en rangeant les caractéristiques entre zéro et 1. Les valeurs de ces deux caractéristiques ont une corrélation directe avec la probabilité d'être une communauté de spam.

TsWSS: Chaque communauté est inférée par un ensemble de tweets posté par ses utilisateurs. Étant donné que les spammeurs automatisent leurs campagnes de spam, la probabilité de trouver une corrélation dans le style d'écriture parmi les tweets considérés est élevée. Par exemple, les tweets spam d'une campagne ont une structure de style commune pour écrire des tweets (mot, mot, hashtag, mot, mot, mot, mot, mot et après URL). Nous modélisons cette caractéristique en transformant d'abord les textes des tweets en un niveau supérieur d'abstraction. Ensuite, nous mesurons la similarité du style d'écriture parmi les tweets de la j^{eme} communauté en utilisant la similarité de Jaccard. Une fonction de transformation $Type(ST) \in \{W, H, U, M\}$, est définie qui prend une chaîne de caractère ST comme paramètre et renvoie le type de la chaîne d'entrée (**W**ord, **H**ashtag, **U**rl et **M**ention). Par conséquent, pour un tweet $TW_\bullet$, le nouvel ensemble de représentation d'abstraction est $Trans(TW_\bullet) = \{(i, Type(S))|(I, S) \in TW_\bullet.Text\}$ où i est la position de la chaîne de mots dans le texte du tweet et S est une chaîne de mots qui nécessite une transformation. Avec ces définitions, nous calculons la similarité de style d'écriture parmi un ensemble de tweets comme suit:

$$TsWSS(C_j) = \frac{\sum_{T_1 \in Tweets_j} \sum_{T_2 \in Tweets_j} \frac{|Trans(T_1) \cap Trans(T_2)|}{|Trans(T_1) \cup Trans(T_2)|}}{(|Tweets_j|)(|Tweets_j| - 1)} \tag{8}$$

où $Tweets_j = \bigcup_{u \in C_j} u.TS$ est une unification de tous les tweets affichés par les utilisateurs de la j^{eme} communauté inférée. Les valeurs supérieures et inférieures de cette fonction sont un et zéro respectivement. La valeur élevée de cette fonction signifie que la probabilité de la j^{eme} communauté d'être un spam est élevée en raison de la proximité des tweets dans le style d'écriture.

TPBS: Le comportement de partage (*posting behaviour*) (par exemple, toutes les 5 min) de tweets au niveau de synchronisation pourrait être un indice important supplémentaire dans l'identification des communautés de spam. Ainsi, nous proposons une caractéristique qui mesure la corrélation entre le comportement de partage des utilisateurs. Nous modélisons ce comportement en calculant d'abord la distribution de probabilité de temps d'affichage de chaque utilisateur appartenant à une communauté particulière.

Pour chaque couple d'utilisateurs, nous mesurons la similarité de leurs distributions de probabilité de temps d'affichage en utilisant le concept de corrélation croisée, ce qui donne une valeur réelle unique comprise entre 0 et 1. Ensuite, nous calculons la distribution de probabilité de la similarité de temps d'affichage pour la comparer avec une distribution uniforme tracée sur l'intervalle [0,1]. Plus formellement, soit P^u_{TS} une distribution de probabilités de temps d'affichage des tweets de l'utilisateur u. P^u_{TS} peut être tiré simplement à partir du temps tweets de l'utilisateur u qui les a déjà affichés dans la rubrique T. Nous calculons la similarité entre deux distributions de temps d'affichage de deux utilisateurs différents appartenant à $u_1, u_2 \in C^{j^{eme}}$ communauté, comme suit:

$$PostSim(u_1, u_2) = \frac{Area(P^{u_1}_{TS} \star P^{u_2}_{TS})}{Min(Area(P^{u_1}_{TS} \star P^{u_1}_{TS}), Area(P^{u_2}_{TS} \star P^{u_2}_{TS}))} \quad (9)$$

Où $Area(\bullet)$ est une fonction qui calcule la zone sous la nouvelle distribution résultante, $Min(\bullet, \bullet)$ est une opération minimale qui sélectionne la zone minimale. Le point clé de la prise de l'opération min est de normaliser la zone qui résulte de la corrélation croisée. La valeur faible de $PostSim$ signifie que les deux utilisateurs d'entrée ont une faible corrélation dans le comportement de temps d'affichage.

En calculant la valeur finale de la fonction $TPBS$, nous calculons d'abord la distribution de probabilité de $PostSim$ sur toutes les paires d'utilisateurs possibles existant dans la $C^{j^{eme}}$ communauté. Soit $P_{PostSim}$ (par exemple $\{(0.25, 0.4), (0.1, 0.6)\}$) la distribution de probabilités de la similarité d'écriture et $P^{Uniform}_{PostSim}$ (par exemple $\{(0.25, 0.5), (0.1, 0.5)\}$) la distribution uniforme correspondante de $PostSim$. Nous quantifions les différences entre les distributions en effectuant une corrélation croisée entre elles, définie comme suit:

$$TPBS(C_j) = 1 - \frac{Area(P_{PostSim} \star P^{Uniform}_{PostSim})}{Area(P^{Uniform}_{PostSim} \star P^{Uniform}_{PostSim})} \quad (10)$$

Où la valeur élevée (près de 1) de $TPBS$ signifie que tous les utilisateurs de la j^{eme} communauté ont presque le même comportement positif (c'est-à-dire presque la même fréquence d'affichage) et que cette communauté a une forte probabilité d'être une campagne de spam. D'une autre côté, lorsque le $P_{PostSim}$ est proche de la distribution uniforme, cela signifie que presque aucun utilisateur n'a le même comportement de publication et donc que la communauté a une faible probabilité d'être une campagne de spam.

3.2.4. Fonction de Classification

Nous utilisons les quatre caractéristiques proposées basées sur la communauté pour prédire l'étiquette de classe de chaque utilisateur qui partage des tweets dans la rubrique T. Nous classons les utilisateurs d'une communauté en spam si et seulement si l'une des quatre caractéristiques est supérieure à un seuil donné. L'intuition derrière cette proposition est que les caractéristiques conçues sont de détecter les communautés de spam, ce qui signifie que d'avoir au moins une valeur de caractéristique élevée est suffisant pour juger si une communauté est spam. Alors que, pour juger une communauté comme non-spam, il est obligatoire de s'assurer que toutes les caractéristiques sont de faible valeur. Par conséquent, nous concevons la fonction de classification finale, y, de sorte qu'elle

attribue un label à un utilisateur d'entrée $u \in C_j$ basé sur les quatre caractéristiques communautaires, définies comme suit:

$$\text{y(u)} = \begin{cases} spam & u \in C_j \;\&\; (TsWss(C_j) \geq \Delta || \; TPBS(C_j) \geq \Delta \\ & \quad || \; SNPS(C_j) \geq \Delta || \; UNPS(C_j) \geq \Delta) \\ non-spam & u \in C_j \;\&\; TsWss(C_j) < \Delta \& \; TPBS(C_j) < \Delta \\ & \quad \& \; SNPS(C_j) < \Delta \& \; UNPS(C_j) < \Delta \end{cases} \tag{11}$$

Où Δ est un seuil de classification déterminé en fonction de la métrique de performance (par exemple exactitude, rappel, précision) qui doit être optimisée.

4. Description de la base de données et de la vérité terrain

Méthode d'aspiration (*Crawling)*. Nous exploitons notre *crawler* de l'équipe de recherche pour collecter des comptes et des tweets, lancés depuis le 1/Jan/2015. l'approche de diffusion en continu est utilisée pour obtenir un accès pour 1% des tweets globaux, en tant qu'approche d'analyse impartiale. Une telle approche est couramment exploitée dans la littérature pour recueillir et créer des données dans les recherches sur les OSN.

Tableau 1. Statistiques des utlisateurs (comptes) et des tweets annotés

	Spam	non-Spam
Number of Tweets	763,555 (11.8%)	5,707,254 (88,2%)
Number of Users	185,843(8.9%)	1,902,288(91.1%)

Description de la base de données. En utilisant notre ensemble de données Twitter de l'équipe, nous avons regroupé les tweets collectés en fonction de la thématique disponible dans le tweet en ignorant les tweets qui ne contiennent aucune thématique. Ensuite, nous avons sélectionné les tweets de 100 thématiques tendances (par exemple #Trump) échantillonnés au hasard pour mener nos expérimentations. Nous exploitons de cette façon l'exploration et l'échantillonnage pour éliminer toute polarisation possible dans les données et tirer des conclusions impartiales.

Base de données vérité terrain. Pour évaluer l'efficacité des caractériqtiques (*patterns)* décrivant le spam dans la récupération des comptes spam, nous avons créé un ensemble de données annotées en étiquetant chaque compte comme spam ou non-spam. Cependant, avec l'énorme quantité de comptes, l'utilisation de l'approche par annotation manuelle pour avoir des jeux de données étiquetés est une solution peu pratique. Par conséquent, nous exploitons un processus d'annotation largement suivi dans les recherches de détection de spam social. Le processus vérifie si l'utilisateur de chaque tweet a été suspendu par Twitter. En cas de suspension, l'utilisateur avec ses tweets sont étiquetés comme spam; sinon nous assignons non-spam pour les deux. Au total, comme indiqué dans le Tableau 1, nous avons constaté que plus de 760 000 tweets ont été classés comme spam, génèrés par près de 185 800 comptes spam.

5. Résultats et Evaluations

Nous présentons dans cette section, les résultats obtenus en comparant notre approche avec d'autres algorithmes de la littérature.

5.1. Configuration expérimentale

Métriques de précision. Comme la vérité terrain de chaque classe d'étiquette de chaque tweet est donnée, nous utilisons l'exactitude (*accuracy)*, la précision, le rappel, la F-mesure, la précision moyenne, le rappel moyen et la F-mesure moyenne; calculée en fonction de la matrice de confusion de l'outil Weka (Hall *et al.*, 2009); comme métriques couramment utilisées dans les problèmes de classification. Comme notre problème est la classification en deux classes (binaires), nous calculons la précision, le rappel et la F-mesure pour la classe «spam», alors que les métriques de moyennes combinent les deux classes en fonction de la fraction de chaque classe (par exemple 11,8 % * "Précision de spam" + 88,2 % * "précision de non-spam").

***Baselines* ou données de référence.** Nous définissons deux baselines pour comparer notre approche avec eux, à savoir: (i) baseline "A" qui représente les résultats lors de la classification de tous les tweets comme non-spam directement sans classement; (ii) baseline "B" qui montre les résultats obtenus lors de l'application d'algorithmes d'apprentissage supervisés selon les caractéristiques associées au "tweet" . Comme de nombreux algorithmes d'apprentissage fournis par l'outil Weka, nous exploitons *Naive Bayes*, *Random Forest*, *J48,* et *Support Vector Machine* (SVM) comme approches d'apprentissage supervisé connues pour évaluer la performance des caractéristiques mentionnées.

Tableau 2. Résultats de performance du baseline A et du baseline B selon différents paramètres.

Learning Algorithm	Accuracy	Precision	Recall	F-Measure	Avg. Precision	Avg.Recall	Avg. F-Measure
Baseline (A): All Tweets Labeled as Non-Spam							
————	91.1%	0.0%	0.0%	0.0.%	91.1%	91.1%	91.1%
Baseline (B): Supervised Machine Learning Approach							
Naive Bayes	81.2%	13.7%	10.5%	11.9%	79.0%	81.2%	80.1%
Random Forest (#Trees=100)	86.4%	13.2%	2.8%	4.6%	79.0%	86.4%	80.1%
Random Forest (#Trees=500)	86.5%	12.6%	2.6%	4.7%	79.4%	86.5%	82.8%
J48 (Confidence Factor=0.2)	86.4%	13.8%	2.9%	4.9%	79.6%	86.4%	82.5%
SVM (Gamma=0.5)	87.2%	15.7%	0.2%	0.4%	78.3%	87.2%	82.5%
SVM (Gamma=1.0)	87.0%	15.9%	0.1%	0.3%	77.9%	87.0%	82.2%

Paramétrage des *Baselines*. Pour l'approche *Naive Bayes*, nous définissons les options "*useKernelEstimator*" et "*useSupervisedDiscretization*" à false comme valeurs par défaut définies par Weka. Pour *Random Forest*, nous avons mis l'option "*max depth*" à 0 (illimité), en étudiant l'effet du changement du nombre d'arbres $\in \{100, 500\}$. Pour l'approche $J48$, nous fixons le nombre minimum d'instances par feuille à 2, le nombre de plis à 3 et le facteur de confiance à 2. Pour l'approche SVM, nous utilisons l'implémentation *LibSVM* (Chang, Lin, 2011) intégrée à l'outil Weka pour définir la fonction du noyau sur *Radial Basis* et examiner l'impact de gamma $\in \{0.5, 1\}$, où les paramètres restants sont par défaut.

Paramétrage de notre approche. Pour l'étape de détection communautaire, nous fixons η = 0.001, M = 10,000 et $\epsilon = 0.0001$ comme valeurs pour le taux d'apprentis-

sage, le nombre d'itérations et le seuil de changement absolu dans la matrice cachée **H**, respectivement. Pour le nombre de communautés K, nous expérimentons notre approche à deux valeurs différentes, $K \in \{5, 10\}$, pour en étudier l'effet. Pour la taille des matrices d'information **X**, nous considérons tous les utilisateurs distincts (comptes) de chaque hashtag sans exclure aucun utilisateur disponible dans la collection d'essai. Comme un algorithme itératif est utilisé pour résoudre le problème d'optimisation de la détection communautaire, nous initialisons chaque entrée de la matrice cachée **H** par une petite valeur réelle positive tirée d'une distribution uniforme sur l'intervalle $[0, 1]$. Pour le seuil Δ, nous étudions l'impact de la modification de sa valeur en effectuant des expérimentations à différentes valeurs de $\Delta \in [0.1, 1.0]$ avec un pas d'incrémentation de 0.1.

5.2. Résultats Expérimentaux

Résultats de *Baselines*. Selon les résultats des baselines rapportées dans le Tableau 2, les modèles de classification supervisés ont une forte défaillance dans le filtrage des tweets spam existant dans les 100 thématiques tendances. Cet échec peut être facilement identifié à partir des valeurs basses de rappel de spam (4^{eme} colonne) où la valeur la plus élevée est obtenue par l'algorithme d'apprentissage $NaiveBayes$. Le 10,5 % du rappel de spam obtenu par $NaiveBayes$ signifie que moins de 80 000 tweets spam peuvent être détectés à partir de 736 500 tweets spam. Les faibles valeurs de précision du spam indiquent également qu'un nombre important de tweets «non-spam» a été classé en «spam». Par la suite, comme la F-mesure de spam dépend des mesures de rappel et de précision, les valeurs de la F-mesure de spam sont évidemment faibles. Les valeurs de précision du baseline «B» sont proches des valeurs de précision de la baseline «A». Toutefois, compte tenu des faibles valeurs de précision du spam et du rappel de spam, la métrique d'exactitude dans ce cas n'est pas une mesure indicative et utile pour juger l'apprentissage supervisé comme une approche efficace. Plus précisément, l'approche d'apprentissage supervisé n'ajoute pas une contribution significative à l'amélioration de la qualité des 100 tweets des thématiques tendances. L'idée clé de l'utilisation de différents algorithmes d'apprentissage est de varier leurs paramètres est de mettre en évidence la mauvaise qualité des techniques de l'état de l'art traitant le tweet. Dans l'ensemble, les résultats obtenus par les modèles permettent de tirer diverses conclusions: (i) les techniques de l'état de l'art ne sont pas discriminatives entre les tweets non-spam et spam, assurant la dynamique du contenu spam; (ii) les spammeurs ont tendance à publier des tweets presque similaires aux non-spam; (iii) l'adoption d'une approche supervisée pour effectuer l'apprentissage sur un ensemble de données annotées de thématiques tendances et l'application du modèle de classification sur des thématiques tendances futures ou non annotées n'est *pas* la solution du tout.

Effet de Δ. En examinant les performances de notre approche dans le Tableau 3, le comportement est complètement différent lors du rappel (classement) des comptes spam, surtout lorsque la valeur de Δ devient plus faible. Les résultats de rappel sont tout à fait compatibles avec l'équation 11 conçue pour classer les utilisateurs. Pour des valeurs faibles de Δ, certaines communautés non-spam sont classées comme spam. En effet, cela explique la dégradation dramatique de la précision lors de la diminution de

Tableau 3. Les résultats de performance de notre approche en plusieurs métriques, calculées à différentes valeurs de nombre de communautés $K \in \{5, 10\}$ et à divers seuils de classification $\Delta \in \{0.1, 1\}$.

Model (Δ)	Accuracy	Spam Precision	Spam Recall	Spam F-measure	Avg. Precision	Avg. Recall	Avg. F-measure
Number of Communities ($K = 5$)							
Δ=0.1	63.0%	14.6%	**65.8%**	23.9%	**87.9%**	63.0%	73.4%
Δ=0.2	66.3%	15.4%	63.0%	24.8%	87.8%	66.3%	75.6%
Δ=0.3	68.1%	15.9%	60.7%	**25.2%**	87.8%	68.1%	76.7%
Δ=0.4	69.7%	16.1%	57.5%	25.1%	87.5%	69.7%	76.7%
Δ=0.5	77.4%	18.2%	44.3%	25.8%	87.0%	77.4%	81.9%
Δ=0.6	78.2%	17.7%	40.3%	24.6%	86.7%	78.2%	82.2%
Δ=0.7	82.0%	18.6%	30.9%	23.3%	86.3%	82.0%	84.0%
Δ=0.8	85.6%	19.7%	20.2%	19.9%	85.8%	85.6%	85.7%
Δ=0.9	89.1%	**20.3%**	7.6%	11.1%	85.2%	89.1%	**87.1%**
Δ=1.0	**91.1%**	0.0%	0.0%	0.0%	83.1%	**91.1%**	86.9%
Number of Communities ($K = 10$)							
Δ=0.1	69.0%	15.9%	**58.6%**	25.1%	87.6%	69.0%	77.2%
Δ=0.2	71.5%	16.7%	56.0%	25.8%	**87.6%**	71.5%	78.8%
Δ=0.3	73.3%	17.2%	53.3%	**26.0%**	87.5%	73.3%	79.7%
Δ=0.4	75.1%	17.5%	49.0%	25.8%	87.2%	75.1%	80.7%
Δ=0.5	81.3%	19.5%	35.7%	25.2%	86.7%	81.3%	83.9%
Δ=0.6	82.1%	19.0%	31.5%	23.7%	86.3%	82.1%	84.2%
Δ=0.7	85.1%	19.8%	22.3%	21.0%	85.9%	85.1%	85.5%
Δ=0.8	87.7%	**20.7%**	13.5%	16.3%	85.6%	87.7%	86.6%
Δ=0.9	90.0%	20.7%	4.0%	6.6%	85.0%	90.0%	**87.4%**
Δ=1.0	**91.1%**	0.0%	0.0%	0.0%	83.1%	**91.1%**	86.9%

la valeur de Δ, ainsi que la dégradation de la précision du spam. Bien que les valeurs de rappel sont élevées, les valeurs de précision de spam de notre approche sont presque semblables à celles de l'approche d'apprentissage supervisé.

Impact de K. Le nombre de communautés K, a un impact direct et évident sur l'exactitude (*accuracy)*, la précision et les mesures de rappel de spam. En effet, l'exactitude et la précision du spam augmentent tant que le nombre de communautés augmente. La justification de ce comportement est que les 100 thématiques tendances expérimentés ont été attaqués par de nombreuses campagnes de spam non corrélées. Par la suite, l'augmentation du nombre de communautés permet de détecter les campagnes de spam de manière précise et exacte. Au niveau du rappel de spam, le comportement est inversement corrélé au nombre de communautés. L'utilisation d'un plus grand nombre de communautés que le nombre réel (inconnu) de campagnes de spam conduit à séparer ces campagnes sur plus de communautés. Comme les communautés de spam peuvent contenir des utilisateurs non-spam (comptes); dans ce cas, les valeurs des caractéristiques diminuent, classifiant ces communautés comme "non-spam".

6. Conclusion

Dans cet article, nous avons décrit une approche non supervisée pour filtrer les utilisateurs spam dans les collections à grande échelle de thématiques "tendances". Notre approche adopte la perspective collective dans la détection des utilisateurs spammeurs en découvrant les corrélations entre eux. Notre travail apporte deux contributions au domaine de la qualité de l'information: (i) le filtrage des utilisateurs sans avoir besoin de jeux de données annotés; (ii) un processus de filtrage rapide en raison de la dépendance des méta-données disponibles, sans avoir recours à l'information des serveurs de Twitter. Avec cette nouvelle idée, nous planifions d'étudier l'impact de la collaboration avec d'autres OSN pour améliorer les résultats actuels.

Remerciements

Ce travail s'intègre dans le cadre des contributions du projet ANR FILTER 2.

Bibliographie

Abascal-Mena R., Lema R., Sèdes F. (2015). Detecting sociosemantic communities by applying social network analysis in tweets. *Social Netw. Analys. Mining*, vol. 5, n° 1, p. 38:1–38:17. Consulté sur http://dx.doi.org/10.1007/s13278-015-0280-2

Benevenuto F., Magno G., Rodrigues T., Almeida V. (2010). Detecting spammers on twitter. In *In collaboration, electronic messaging, anti-abuse and spam conference (ceas)*, p. 12.

Canut C. M., On-at S., Péninou A., Sèdes F. (2015). Time-aware egocentric network-based user profiling. In *Proceedings of the 2015 IEEE/ACM international conference on advances in social networks analysis and mining, ASONAM 2015, paris, france, august 25 - 28, 2015*, p. 569–572. Consulté sur http://doi.acm.org/10.1145/2808797.2809415

Cao C., Caverlee J. (2015). Detecting spam urls in social media via behavioral analysis. In *Advances in information retrieval*, p. 703–714. Springer.

Chang C.-C., Lin C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, vol. 2, p. 27:1–27:27. (Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm)

Chu Z., Gianvecchio S., Wang H., Jajodia S. (2012). Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *Dependable and Secure Computing, IEEE Transactions on*, vol. 9, n° 6, p. 811–824.

Chu Z., Widjaja I., Wang H. (2012). Detecting social spam campaigns on twitter. In *Applied cryptography and network security*, p. 455–472.

Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I. H. (2009, novembre). The weka data mining software: An update. *SIGKDD Explor. Newsl.*, vol. 11, n° 1, p. 10–18.

Martinez-Romo J., Araujo L. (2013). Detecting malicious tweets in trending topics using a statistical analysis of language. *Expert Systems with Applications*, vol. 40, n° 8, p. 2992–3000.

McCord M., Chuah M. (2011). Spam detection on twitter using traditional classifiers. In *Proceedings of the 8th international conference on autonomic and trusted computing*, p. 175–186. Springer-Verlag.

Stringhini G., Kruegel C., Vigna G. (2010). Detecting spammers on social networks. In *Proceedings of the 26th annual computer security applications conference*, p. 1–9. New York, NY, USA, ACM.

Wang A. H. (2010, July). Don't follow me: Spam detection in twitter. In *Security and cryptography (secrypt), proceedings of the 2010 international conference on*, p. 1-10.

Yang J., Leskovec J. (2013). Overlapping community detection at scale: A nonnegative matrix factorization approach. In *Proceedings of the sixth acm international conference on web search and data mining*, p. 587–596. New York, NY, USA, ACM. Consulté sur http://doi.acm.org/10.1145/2433396.2433471

Approche temporelle pour la génération personnalisée de profils folksonomiques

Tahar-Rafik Boudiba [1] [2], **Rachid Ahmed-Ouamer** [1]

1. *LARI laboratory Department of computer science*
   Mouloud Mammeri University 15000, Tizi-Ouzou, Algeria.
   Rachid.ahmedouamer@yahoo.fr
2. *IRIT IRIS, 118 Route de Narbonne*
   31062, Toulouse CEDEX 9 France.
   Tahar-Rafik.Boudiba@irit.fr

RÉSUMÉ. *L'annotation collaborative offre aux utilisateurs la possibilité de décrire des ressources avec des mots-clés (tags). Un tag décrit l'intérêt que l'utilisateur porte à la ressource. L'ensemble des triplets {utilisateur, ressource, tags} constitue une folksonomie. Dans cet article, nous proposons de construire des profils utilisateurs à partir des tags. Ces profils (folksonomiques) sont des clusters contenant des ressources correspondant aux divers centres d'intérêts de l'utilisateur. Comme les intérêts utilisateurs évoluent avec le temps, nous estimons qu'en pondérant les tags en fonction non seulement de leur qualité supposée, mais aussi de leur "fraîcheur", on améliore le classement des ressources. Pour évaluer la "pertinence" de ces profils, nous les avons utilisés dans un contexte de recherche d'information personnalisée. En comparant notre approche avec une approche de base (sans facteur temporel), nous avons mené une série d'expérimentations sur des données extraites à partir de MovieLens. Les résultats confirment l'efficacité de notre approche.*

ABSTRACT. *Collaborative annotation offers users the possibility to describe resources with keywords (tags). A tag describes the user's interest to the resource. The set of triplets {user, resource, tags} constitutes a folksonomy. In this paper, we propose to construct user profiles from tags. These (folksonomy based) user profiles are clusters containing classified resources corresponding to the various interests of the user. As user interests evolve over time, we estimate that by weighting the tags according to not only their supposed quality but also their "freshness", the classification of resources is improved. To evaluate the "relevance" of these profiles, we used them in a personalized information retrieval context. By comparing our approach with a basic one (without time factor), we conducted a series of experimentations extracted from MovieLens Dataset. The results confirm the effectiveness of our approach.*

MOTS-CLÉS : *Folksonomies, recherche d'information, temporalité, profil utilisateur, clustering.*

KEYWORDS: *Folksonomies, information retrieval, temporality, user profile, clustering.*

1. Introduction

Avec le développement fulgurant que connaˆInternet, à travers les systèmes d'annotations collaboratives, le Web 2.0 est devenu de plus en plus populaire, ce qui a fait passer les utilisateurs de consommateurs passifs en producteurs actifs de contenus Web (Zanardi, Capra, 2008).

Les usagers de ces systèmes génèrent une énorme quantité de données personnelles. Ils utilisent des mots-clés ou tags leur permettant de catégoriser, d'indexer, ou de classer diverses ressources qui concordent avec les intérêts de ces usagers. L'ensemble des triplets { utilisateur, ressource, tags } constitue une folksonomie.

Les tags décrivent les intérêts que portent les utilisateurs aux ressources. Ils fournissent des informations utiles à la construction de profils utilisateurs. En général, ces profils sont représentés sous forme de vecteurs de termes (tags) auxquels on associe un poids w à chaque terme t . Certains travaux (Xu *et al.*, 2008 ; Vallet *et al.*, 2010), utilisent *tf-idf* ainsi que des variantes de ce modèle comme schéma de pondération.

En partant du constat que l'ensemble des tags utilisateurs décrivent les préférences de celui-ci à l'égard d'une ressource, la dimension temporelle de l'annotation fournit quant à elle, une indication sur l'évolution de l'intérêt de l'utilisateur à travers le temps (Zheng, Li, 2011). Certains travaux intègrent cette dimension lors de la construction de profils utilisateurs (Kacem *et al.*, 2014 ; Cheng *et al.*, 2008 ; Abel *et al.*, 2011). De manière générale, la temporalité associée à l'action d'annotation est quantifiée à l'aide d'une stratégie de pondération visant à attribuer un poids à chaque ressource en se basant sur les informations que fournissent les tags (popularité ou timestamps) ou une combinaison des deux (Zheng, Li, 2011).

Cependant les travaux exploitant ce type d'information ne nous renseignent pas à quel moment un contenu devient obsolète. De plus, si certains utilisateurs actifs annotent des ressources, et cela, de manière plus fréquente que d'autres utilisateurs moins actifs, alors les poids du tag de l'utilisateur le plus actif seront plus élevés que ceux du moins actif. En ce qui concerne les ressources, les utilisateurs actifs influencent la fréquence avec laquelle elles ont été annotées, les rendant plus populaires du fait de la tendance disproportionnée des utilisateurs actifs à annoter ces ressources. Cela affecte de ce fait le classement global de celles-ci, qui subissent une distorsion systématique causée par des utilisateurs plus actifs que d'autres.

Dans cet article nous proposons de construire des profils qui prennent en considération l'aspect temporel des annotations, nous regroupons en catégorie les ressources en intégrant une formule de pondération normalisée modérée par une fonction temporelle. Ces profils (folksonomiques) sont des clusters contenant des ressources classées correspondant aux divers intérêts de l'utilisateur ainsi que de leurs évolutions dans le temps. Nous évaluons par la suite notre approche en la comparant avec une approche

de base sur une collection de données extraites à partir de MovieLens [1] qui confirment l'efficacité de notre approche.

L'article est structuré de la manière suivante: nous présentons dans la section 2 un aperçu de certains travaux connexes, après quoi nous détaillons notre approche dans la section 3. La section 4 est consacrée à l'expérimentation. Enfin, la section 5 conclut l'article et énonce certaines perspectives.

2. État de l'art

Dans cette section, nous présentons certaines approches de personnalisation exploitant les folksonomies pour la construction de profils, nous nous intéresserons aussi aux travaux intégrants la temporalité pour la construction de profils utilisateurs ou visant à prendre en compte l'évolution de ces derniers. Nous faisons une synthèse de quelques-unes de ces approches en les présentant en deux sous-sections, l'une traitant des approches de personnalisation basée sur les folksonomies, tandis que l'autre portera sur les approches temporelles se basant sur les tags.

2.1. Approche de personalisation basé sur les folksonomies

Les approches décrites ci-dessous utilisent le modèle vectoriel pour la génération de profils. Dans (Xu *et al.*, 2008), les auteurs utilisent la mesure du cosinus afin de calculer la similarité entre profils utilisateurs et profils de documents. L'approche est basée sur le modèle vectoriel. Comme schéma de pondération (Xu *et al.*, 2008) utilisent *tf-idf* qui est basé sur le modèle *BM25*. Noll et Meinel (Noll, Meinel, 2007) supposent que les tags plus fréquemment utilisés sont plus intéressants et plus pertinents pour un utilisateur que ceux utilisés rarement, l'idée est de tirer parti des annotations fournies par la communauté pour identifier les perceptions communément admises des documents. (Vallet *et al.*, 2010) présentent une approche de personnalisation avec comme schéma de pondération *tf-idf*. Celle-ci est similaire à celle de (Xu *et al.*, 2008), cependant Vallet et al. éliminent le facteur de normalisation de la longueur du document. Dans le modèle vectoriel le but d'une telle manœuvre est de pénaliser le score des documents qui contiennent un grand nombre d'informations, et qui de ce fait sont susceptibles de correspondre à la requête par hasard. En ce qui concerne les systèmes d'annotations sociales, le fait qu'un grand nombre de tags soit connexes est lié à la popularité des ressources. Selon (Cai, Li, 2010) les schémas *tf*, *tf-idf* ou *BM25* utilisés pour pondérer des tags, se révèlent être insuffisant pour indiquer à quel point un utilisateur présente de l'intérêt pour les tags décrivant son profil. De plus, des utilisateurs actifs influencent les fréquences avec lesquelles les ressources ont été annotées, les rendant plus populaire du fait de la tendance disproportionnée des utilisateurs actifs à annoter ces ressources. Cela affecte de ce fait le ranking (classement) des ressources à la recherche.

1. http:/movielensorg/

2.2. Approches temporelles pour la construction de profils utilisateurs

La motivation principale de ce type d'approche réside dans le fait que les intérêts des utilisateurs évoluent avec le temps. En effet ils déclinent au fur et à mesure que le temps passe (Zheng, Li, 2011). Les travaux décrivant ces approches (Abel *et al.*, 2011 ; Zheng, Li, 2011 ; Kacem *et al.*, 2014) exploitent des fonctions temporelles qui simulent l'atténuation des intérêts utilisateurs au cours du temps.

Dans (Zheng, Li, 2011), les auteurs développent une fonction exponentielle adaptative qui simule l'atténuation graduelle des intérêts utilisateurs. Zheng et al. combinent deux mesures de poids; l'une concerne la ressource par rapport au tag qui le décrit, l'autre fait état de la valeur temporelle attribuée à la ressource.

Dans (Kacem *et al.*, 2014), le profil utilisateur est représenté à l'aide d'un vecteur de termes. L'importance de chaque terme est ajusté en fonction de sa date d'apparition. Cette mesure de "fraîcheur" de terme est calculée grâce à une fonction à noyaux gaussiens, qui permet le calcul d'une distance temporelle entre la date courante et la date ou apparaˆle terme (l'interaction sociale). Abel et al. (Abel *et al.*, 2011) Analysent les activités individuelles sur Twitter [2]. Ils comparent différentes stratégies pour la création de profils utilisateurs, étudient comment ces profils changent au fil du temps, et comment la dynamique temporelle influence l'exactitude du processus de personnalisation. Les auteurs considèrent deux types de profil utilisateur: les utilisateurs actifs qui interagissent pendant une longue période de temps, et ceux qui sont moins actifs et qui interagissent de manière plus irrégulière.

3. Approche temporelle pour la génération de profils folksonomiques

3.1. Notations

Une folksonomie *F* peut être définie comme étant un 4-uplet F= (U, T, R, A), où U est l'ensemble des utilisateurs annotant les ressources de l'ensemble *R* avec $U = \{u_1, u_2, ...u_m\}$ où chaque u_i est un utilisateur; *T* est l'ensemble des tags que comprend le vocabulaire exprimé par la folksonomie $T = \{t_1, t_2, ...t_l\}$; *R* est l'ensemble des ressources annotées par les utilisateurs $R = \{r_1, r_2...r_n\}$; $A = \{u_m, t_l, r_n\} \in U \times T \times R$ est l'ensemble des annotations de chaque tag t_l à une ressource r_n par un utilisateur u_m. Le profil utilisateur u_m peut être défini comme un vecteur $\vec{u}_m = (u_{m,1}, ...u_{m,l})$, où $u_{m,l} = |\{(u_m, t_l, r) \in A / r \in R\}|$ est le nombre de fois que l'utilisateur a annoté une ressource avec le tag t_l.

3.2. Description générale de l'approche

On considère un système non personnalisé *S* retournant un ensemble de ressources *R*, et qui fournit une liste de ressources classées $S(q) \subseteq R$. On suppose que $\forall q$, $S(q)$

2. https://twitter.com/

satisfait une requête q de l'utilisateur. Le classement suit l'ordre implémenté par le système $[S = r_1 \geq r_2 \geq ...r_k]$ où $r_k \in R$.

Notre approche consiste à adopter une formule de pondération normalisée modérée par une fonction temporelle afin d'attribuer un poids à chaque tag. Ce poids sera représentatif de la qualité supposée du tag à décrire la ressource, mais aussi de la "fraîcheur", de celui-ci. Dans le but de proposer un contenu personnalisé à l'utilisateur, nous catégorisons les ressources annotées par celui-ci via un clustering ; les ressources seront regroupées en catégories d'intérêts $C = (C_1, C_2...C_h)$. Enfin, nous définissons (section 4.2) une signature propre à chaque cluster comme étant la proportion de tags qui apparaˆle plus dans le cluster.

Nous considérons les signatures des clusters comme des requêtes représentées par des vecteurs de tags $\vec{q} = (t_1, t_2, ...t_n)$. Nous les utiliserons pour récupérer les ressources en comparant les tags des signatures à l'ensemble des tags décrivant les ressources contenues dans le cluster. Afin de classer ces ressources par similarités, nous calculons pour chaque ressource appartenant à C représentée comme un vecteur de tags $\vec{e} = (t_1, t_2, ...t_m)$ la similarité entre les deux vecteurs $\overrightarrow{e}$ et $\overrightarrow{q}$, cette approche personnalisée fournit une liste de ressources classées $S'(q, e) \subseteq R$, en accord avec les préférences de l'utilisateur *u*. Formellement $[S' = r_1 \geq r_2 \geq ...r_k]$, en définissant la relation d'ordre $\geq$ centrée sur l'utilisateur *u*, par la condition:

$$r_i \geq r_j \geq \Leftrightarrow sim(u, r_i, \overrightarrow{q}) \geq sim(u, r_j, \overrightarrow{q})$$

Où *sim(u,r,* $\overrightarrow{q}$*)* est la fonction de similarité entre *u* et *r* en prenant en considération le classement de *r* dans *S(q)*.

3.3. Schéma de Pondération

Nous proposons de représenter initialement le profil utilisateur sous forme vectorielle. Nous attribuons à chaque tag une valeur $v_{i,k}$ qui correspond à une fréquence de tag normalisée. $\overrightarrow{U_i}$ est un vecteur de tags: valeurs, i.e

$$\overrightarrow{U_i} = (t_{i,1} : v_{i,1}, t_{i,2} : v_{i,2}...t_{i,n} : v_{i,n})$$

Où $t_{i,k}$ est un tag annoté par u_i, *n* est le nombre total de tags annotés par u_i. $v_{i,k}$ est le degré de préférence de l'utilisateur u_i au tag $t_{i,k}$, qui peut être intuitivement obtenu comme suit:

$$v_{i,k} = \frac{N_{i,k}}{N_i} \tag{1}$$

Où $N_{i,x}$ est le nombre de fois qu'un utilisateur *i* annote avec le tag *x*, et N_i est le nombre de ressources annotées par l'utilisateur *i*.

En tenant compte du fait que les préférences des utilisateurs évoluent avec le temps, nous proposons de revisiter la notion de fréquence de tag normalisée en biaisant cette dernière par une fonction temporelle. Nous utilisons une fonction à noyaux

Gaussien (Kacem *et al.*, 2014 ; Badache, Boughanem, 2015) afin d'estimer la distance entre la date courante et la date à laquelle la ressource a été annotée:

$$K(S^c, S_j) = \frac{1}{\sigma\sqrt{2\pi}}.exp\left[\frac{-(S^c - S_j)^2}{2\sigma^2}\right] \quad (2)$$

Où σ est le coefficient d'interpolation, S^c est la date courante, S_j est la date qui correspond à l'action d'annotation.

Par conséquent, nous revisitons la notion de profil utilisateur, en attribuant les pondérations suivantes au vecteur de tags:

$$\overrightarrow{U} = (t_1^{S_j} : W_1^{S_j},\ t_2^{S_j} : W_2^{S_j}, \ldots,\ t_m^{S_j} : W_m^{S_j})$$

Où le poids $W(t)^{S_j}$ d'un tag *t* dans un profil utilisateur est la somme de ses fréquences normalisées et biaisées par la fonction temporelle, et est définie comme suit:

$$W(t_k)^{S_c} = \sum v_{i,k}(t_k)^{S_j}\ K(S^c, S_j) \quad (3)$$

3.4. Clustering

Afin de regrouper les ressources en catégories d'intérêts, nous construisons d'abord un graphe biparti représentant l'assignation des tags utilisateurs à l'ensemble des ressources utilisateurs. Nous attribuons à chaque tag du graphe biparti la pondération normalisée et biaisée par la fonction temporelle $W(t_k)^{Sc}$. Par la suite, en faisant une projection 1-mode sur l'ensemble des ressources, nous obtenons un graphe de ressources où chaque nœud correspond à une ressource. Dès lors, nous pondérons les arêtes entre les ressources par la somme des pondérations $W(t_k)^{Sc}$ des tags en commun reliant les ressources entre elles.

3.5. Description de la collection de test

MoviesLens est un système de recommandation de films où les utilisateurs peuvent attribuer des tags et des notes aux films en fonction de leurs préférences. La collection de test que nous avons extrait, contient des informations sur l'identifiant de l'utilisateur, le tag, ainsi que le film annoté. Elle contient aussi la date à laquelle l'utilisateur a annoté le film (exprimé en format Timestamp). De plus elle fournit l'appréciation (rating) de chaque utilisateur à propos d'un film. Cette "note" attribuée par l'utilisateur évalue le film de 1 à 5 étoiles selon l'appréciation de celui-ci.

La collection contient en outre 2000063 notes(rating) et 465564 tags et 27278 films et a été généré le 31 mars 2015. La collection de données MovieLens a été publiée par le groupe de recherche GroupLens à l'Université de Minisota. Afin d'évaluer notre approche, nous donnons un bref apperçu sur l'ensemble de données. Le **Tableau 1** y présente les principaux détails.

Tableau 1 – *Statistiques sur la collection*

Nombre d'utilisateurs	138493
Nombre de films	27278
TAS(Tag assignement)	465564
Rating	2000063
Periode	January 09,1995 - March 31, 2015

3.6. Expérimentation

à partir de cette collection de données, nous extrayons 100 utilisateurs ayant au moins 100 tags dans leurs activités d'annotations. Puis nous évaluons la performance de l'approche personnalisée en fonction de la pertinence graduée des ressources pour chaque utilisateur et la comparons à celle d'une approche qui ne tient pas compte du paramètre temporel.

En ce qui concerne le regroupement des ressources en communautés, nous utilisons un algorithme de clustering dit glouton *"gready algorithm"*. Cet algorithme de clustering est basé sur une optimisation de la modularité, en effet la mesure quantitative qui garantit la qualité des clusters est la modularité (Newman, 2004 ; 2006). La modularité d'un clustering est le nombre d'arêtes intraclusters, comparé au nombre moyen d'arêtes intraclusters que l'on aurait après brassage aléatoire du graphe.

Par la suite, nous comparons notre méthode de génération de profils utilisateurs, à deux autres approches. La première (Yeung *et al.*, 2008) **Baseline 1**, construit des profils utilisateurs en regroupant les ressources sans attribuer préalablement de poids aux tags. La deuxième, **Baseline 2**, regroupe les ressources en sommant les degrés de préférence de l'utilisateur pour chaque tag en commun entre deux ressources i.e: la somme des $v_{i,k}$ en commun entre deux ressources.

En outre, **Baseline 2** adopte un schéma de pondération normalisée des termes, le but d'une telle manœuvre est de montrer l'impact séparé de la pondération des tags et l'atténuation au cours du temps. Yeung et.al utilisent aussi un algorithme de clustering pour regrouper les ressources, mais ne classent pas les ressources selon les intérêts de l'utilisateur. De plus (Yeung *et al.*, 2008) ne prennent pas en compte l'aspect temporel des annotations.

Ci-aprés un aperçu du nombre de clusters par utilisateur que génère notre approche (figure 1).

Une fois les clusters générés, nous définissons une signature propre à chaque cluster comme étant la proportion de tags qui apparaˆle plus dans celui-ci. La signature fournit un aperçu du regroupement en cluster des ressources, mais aussi un moyen d'identifier chaque cluster grâce à un ensemble de tags qui lui est propre. Nous fixons cette proportion à ρ=40%. les similarités seront calculées entre chaque ressource et classées. Ce classement des ressources correspondra aux préférences de l'utilisateur à travers le temps. à l'issue des expérimentations, nous fixons aussi la valeur du coeffi-

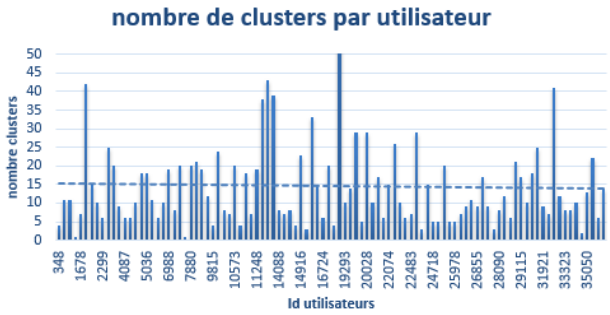


FIGURE 1 – Un aperçu du nombre de clusters générés par utilisateur

cient d'interpolation σ à 4 (l'augmentation de la valeur du paramétre σ réduit sensiblement l'amplitude de la fonction gaussienne 2 section 3.3).

3.7. Évaluation

Les *ratings* (notes) fournis par l'utilisateur, nous apportent un moyen utile de collecter les ressources que l'utilisateur a appréciées. Ils seront utilisés comme jugement de pertinence. Dans le but d'évaluer la qualité du classement des ressources, nous définissons la fonction *"rating"* $f_{u_i}(m_i)$ qui retourne la valeur de la note du film m_i attribuée par l'utilisateur u_i.

La valeur maximale de la fonction est : $Max(F_{u_i}(m_i)) = 5$, cela indique une très bonne appréciation de l'utilisateur à l'égard du film, la valeur minimale de la fonction est quant à elle égale à 1.

Pour une requête donnée, notre système doit fournir une liste classée de résultats, les ressources les mieux classées doivent correspondre aux préférences de l'utilisateur. Dans notre cas pour chaque signature de cluster obtenue grâce à notre approche, une liste de films classés est générée.

Le principe du DCG (Discount Cumulative gain) stipule que les films très pertinents apparaissant au bas d'un résultat de recherche doivent être pénalisés, car la valeur de pertinence graduée est réduite de manière logarithmique et proportionnelle à la position du résultat. Plus la position d'un film pertinent est basse, moins il sera utile pour l'utilisateur, car il est moins susceptible d'être examiné.

Le score de pertinence de chaque film est utilisé comme mesure de la valeur acquise pour sa position classée dans le résultat et le gain est additionné à la position classée de 1 à n. La liste des films classés devient une liste de valeurs acquises en remplaçant les ID du film par leur score de pertinence.

Les scores de pertinence sont définis sur une échelle de 5 étoiles: de 0 à 5, 0 pour des films non notés, 1 à 3 pour des films non pertinents et 4 à 5 pour des films pertinents. Nous présentons dans ce qui suit, le résultat de notre évaluation grâce au calcul du nDCG (Normalized Discounted Cumulative Gain (Jarvelin, Kekalainen, 2002)) (**Tableau 2**). Le DCG idéal étant les films qui sont classés d'après les notes de l'utilisateur.

3.8. Résultats et analyse

On observe (**Figure 2**) une nette amélioration de la position de classement en faveur de notre approche à r@ = 5 à r@ = 10 **Tableau 2**. La valeur nDCG de cette dernière augmente pour atteindre 0.69 à r@ = 90 et r@ = 100, tandis que la **baseline 1** atteint la valeur maximale de 0,62 lorsque r@ = 90 et r@ = 100. La valeur nDCG de la **baseline 2** indique quant à elle une amélioration qui surpasse celle de la **Baseline 1** et atteint une valeur maximale de 0.64 lorsque r@ = 90 et r@ = 100.

De notre point de vue, cela met en évidence d'une part, l'impact de l'attribution de poids qualitatifs représentant les préférences de l'utilisateur, mais aussi l'importance de prendre en compte l'évolution des intérêts de celui-ci. Par conséquent, la combinaison d'un schéma de pondération normalisée de termes avec une mesure permettant l'atténuation progressive de l'importance de ces termes au cours du temps, améliore le classement personnalisé des ressources retournées à l'utilisateur lors du processus de recherche.

Tableau 2 – *Comparaison NDCG*

Rank	NDCG_baseline1	NDCG_baseline2	NDCG_Our
r@5	0.426	0.45	0.56
r@10	0.446	0.5	0.57
r@20	0.465	0.48	0.586
r@30	0.49	0.53	0.587
r@40	0.51	0.55	0.61
r@50	0.53	0.591	0.624
r@60	0.55	0.59	0.63
r@70	0.57	0.611	0.6423
r@80	0.59	0.612	0.67
r@90	0.61	0.63	0.685
r@100	0.62	0.64	0.698

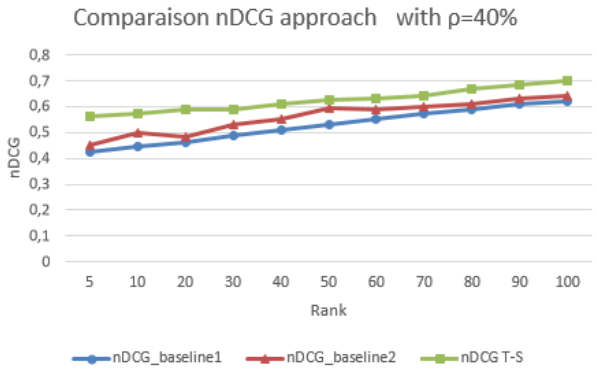


FIGURE 2 – Comparaison nDCG

4. Conclusion

Nous avons proposé une approche visant à générer des profils folksonomiques. Ces profils sont des clusters correspondant aux intérêts de l'utilisateur en tenant compte aussi de l'évolution de ses préférences. Nous avons mis en place une formule de pondération qui prend en compte les aspects qualitatifs et temporels d'une folksonomie. Le modèle de pondération utilisé a été défini grâce à une méthode de clustering. Cette méthode nous a permis d'extraire des ressources classées selon l'ordre de préférence de l'utilisateur et son évolution dans le temps.

En perspective, nous pensons concevoir une étude plus approfondie du paramètre ρ, dans cette optique, des expérimentations sont en cours de validation afin de déterminer avec exactitude la proportion de tags apparaissant le plus dans les clusters générés, et de ce fait, fournir une meilleure description de ces clusters.

Remerciements

Les expériences présentées dans cet article ont été réalisées en utilisant la plateforme OSIRIM qui est administrée par l'IRIT et soutenue par CNRS, la région Midi-Pyrénées, le gouvernement français, et le FEDER (voir http://osirim.irit.fr/site/fr).

Bibliographie

Abel F., Gao Q., Houben G.-J., Tao K. (2011). Analyzing temporal dynamics in twitter profiles for personalized recommendations in the social web. In *Proceedings of the 3rd international web science conference*, p. 2.

Badache I., Boughanem M. (2015). Document priors based on time-sensitive social signals. In *European conference on information retrieval*, p. 617–622.

Cai Y., Li Q. (2010). Personalized search by tag-based user profile and resource profile in collaborative tagging systems. In *Proceedings of the 19th acm international conference on information and knowledge management*, p. 969–978.

Cheng Y., Qiu G., Bu J., Liu K., Han Y., Wang. (2008). Model bloggers' interests based on forgetting mechanism. In *Proceedings of the 17th international conference on world wide web*, p. 1129–1130.

Jarvelin K., Kekalainen J. (2002). Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, vol. 20, n° 4, p. 422–446.

Kacem A., Boughanem M., Faiz R. (2014). Time-sensitive user profile for optimizing search personlization. In *International conference on user modeling, adaptation, and personalization*, p. 111–121.

Newman M. E. (2004). Fast algorithm for detecting community structure in networks. *Physical review E*, vol. 69, n° 6, p. 066133.

Newman M. E. (2006). Modularity and community structure in networks. *Proceedings of the national academy of sciences*, vol. 103, n° 23, p. 8577–8582.

Noll M. G., Meinel C. (2007). Web search personalization via social bookmarking and tagging. In *The semantic web*, p. 367–380. Springer.

Vallet D., Cantador I., Jose J. M. (2010). Personalizing web search with folksonomy-based user and document profiles. In *European conference on information retrieval*, p. 420–431.

Xu S., Bao S., Fei B., Su Z., Yu Y. (2008). Exploring folksonomy for personalized search. In *Proceedings of the 31st annual international acm sigir conference on research and development in information retrieval*, p. 155–162.

Yeung A., Man C., Gibbins N., Shadbolt N. (2008). A study of user profile generation from folksonomies.

Zanardi V., Capra L. (2008). Social ranking: uncovering relevant content using tag-based recommender systems. In *Proceedings of the 2008 acm conference on recommender systems*, p. 51–58.

Zheng N., Li Q. (2011). A recommender system based on tag and time information for social tagging systems. *Expert Systems with Applications*, vol. 38, n° 4, p. 4575–4587.

Gestion de données complexes

Traitement coopératif des requêtes RDF dans le contexte des bases de connaissances incertaines

Ibrahim Dellal, Stéphane Jean, Allel Hadjali, Brice Chardin, Mickaël Baron

LIAS/ISAE-ENSMA - Université de Poitiers
1 Avenue Clement Ader, 86960 Futuroscope Cedex, France
prenom.nom@ensma.fr

RÉSUMÉ. *De nombreuses larges* bases de connaissances *(BC)* incertaines *sont disponibles sur le Web où les faits sont associés avec un degré de confiance α. En général, Les utilisateurs de ces BC n'ayant qu'une connaissance partielle de leur contenu, certaines de leurs requêtes échouent, c'est à dire qu'elles ne renvoient aucun résultat. Pour éviter cette situation frustrante, et au lieu de renvoyer un ensemble vide de réponses, nous avons proposé une approche expliquant l'échec (pour un degré α et pour plusieurs degrés) en fournissant un ensemble de* αMinimal Failing Subqueries *(αMFS), et en calculant des requêtes alternatives, appelées* αMaXimal Succeeding Subqueries *(αXSS), qui sont aussi proches que possible de la requête initiale. Les expérimentations menées sur le benchmark WatDiv montrent l'intérêt de nos approches par rapport à une méthode de référence.*

ABSTRACT. *Several large uncertain Knowledge Bases (KBs) are available on the Web where facts are associated with a certainty degree α. Usually, users have only a partial knowledge of the KBs contents, their queries may be failing i.e., they return no result for the desired certainty. To prevent this frustrating situation, instead of returning an empty set of answers, our approach explains the reasons of the failure (for a single degree α and for several degrees) with a set of αMinimal Failing Subqueries (αMFSs), and computes alternative relaxed queries, called αMaXimal Succeeding Subqueries (αXSSs), that are as close as possible to the initial failing query. The conducted experiments on the WatDiv benchmark show the relevance of our approaches compared to a baseline method.*

MOTS-CLÉS : *BC incertaines, Requêtes SPARQL, Réponse vide.*

KEYWORDS: *Uncertain KB, SPARQL queries, Empty answers.*

1. Introduction

Une *Base de Connaissance* (BC) est un ensemble d'entités (nommées) et de faits sur ces entités. Les techniques récentes d'extraction d'information ont conduit à la construction de grandes BC à partir du Web, comme DBpedia ("DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia", 2015) et Knowledge Vault ("Knowledge Vault: A Web-scale Approach to Probabilistic Knowledge Fusion", 2014). Ces BC contiennent des milliards de faits représentés sous forme de triplets RDF *(sujet, prédicat, objet)* et sont interrogés avec le langage SPARQL. Comme ces BC ont été construites à partir de sources externes, leurs faits peuvent être *incertains* (c'est-à-dire potentiellement incohérents). Pour prendre en compte cette incertitude, des extensions de RDF et de SPARQL ont été proposées afin de supporter des données pondérées par une confiance (Hartig, 2009 ; Tomaszuk *et al.*, 2013). ces travaux associent ainsi un degré de confiance explicite aux faits de la BC et aux résultats des requêtes SPARQL.

Lors de l'interrogation des BC incertaines, les utilisateurs s'attendent à obtenir des résultats de qualité, c'est-à-dire des résultats possédant un degré de confiance supérieur à un seuil donné α. Cependant, comme ces utilisateurs connaissent rarement la structure et le contenu d'une BC, ils peuvent formuler des requêtes trop restrictives et ainsi être confrontés au problème de réponse vide, c'est-à-dire qu'ils n'obtiennent aucun résultat. Une étude réalisée par Saleem et al. sur *les points d'entrée SPARQL* montre que 10% des requêtes soumises à DBpedia entre mai et juillet 2010 retournaient des résultats vides (Saleem *et al.*, 2015). Au lieu de retourner à l'utilisateur un ensemble vide comme réponse à sa requête, le système peut l'aider à comprendre les raisons de cet échec. Une de ces approches est basée sur l'identification des sous-requêtes qui échouent (MFS pour *Minimal Failing Subqueries*), et fourni également un ensemble de sous-requêtes possédant des résultats (XSS pour *Maximal Succeding Subqueries*) (Fokou *et al.*, 2017). Dans cet article, nous nous intéressons à la généralisation de la notion de MFS et de XSS dans le contexte des BC incertaines.

Cet article est organisé comme suit. La section 2 fournit quelques notions de base et formalise le problème abordé. La section 3 motive notre contribution avec un exemple de requête sur une BC incertaine. La section 4 définit les conditions auxquelles nos travaux précédents peuvent être directement adaptés pour trouver les αMFS et αXSS d'une requête RDF. La section 5 décrit deux approches proposées pour calculer les αMFS et les αXSS. La section 6 montre la mise en œuvre et l'évaluation expérimentale réalisée. La section 7 détaille les travaux connexes. Enfin, nous concluons et fournissons quelques perspectives dans la section 8.

2. Préliminaires et Problématique

Nous présentons ici les notions (selon les notations de Pérez et al. (Pérez *et al.*, 2009)) nécessaires à la lecture de l'article et le modèle de confiance utilisé (selon Hartig (Hartig, 2009)).

Modèle de données. Un *triplet RDF* est un triplet (sujet, prédicat, objet) $\in (U \cup B) \times U \times (U \cup B \cup L)$ où, U est un ensemble d'URI, B est un ensemble de ressources anonymes et L est un ensemble de littéraux. Nous notons T l'union $U \cup B \cup L$. Une *base de données RDF* contient un ensemble de triplets RDF (indiqués par T_{RDF}). Chaque triplet RDF est associe à un *degré de confiance* (ou un *degré de certitude*) représentant la fiabilité. Ce degré est associé au triplet par une fonction $tv : T_{RDF} \rightarrow [0, 1]$.

Requêtes RDF. Un *patron de triplet RDF* t est un triplet (sujet, prédicat, object) $\in (U \cup V) \times (U \cup V) \times (U \cup V \cup L)$, où V est un ensemble de variables disjoint de U, B et L. Nous notons $var(t) \subseteq V$ l'ensemble des variables de t. Une *requêtes RDF* est vue comme une conjonction de patrons de triplet: $Q = t_1 \wedge \cdots \wedge t_n$. Le nombre de patrons de triplet d'une requête Q est noté $|Q|$ et les variables de cette dernière sont notés $var(Q) = \bigcup var(t_i)$.

Évaluation d'une requête RDF. Un *mapping* est une fonction partielle $\mu : V \rightarrow T$. Pour un patron de triplets t, nous notons $\mu(t)$ le triplet obtenu en remplaçant les variables de t $var(t)$ par leurs mapping $\mu(var(t))$. Le domaine de μ, $dom(\mu)$, est un sous-ensemble de V où μ est défini. Deux mappings μ_1 et μ_2 sont *compatibles* lorsque pour tout $x \in dom(\mu_1) \cap dom(\mu_2)$, on a $\mu_1(x) = \mu_2(x)$, c'est-à-dire lorsque $\mu_1 \cup \mu_2$ est aussi un mapping. Soit Ω_1 et Ω_2 des ensembles de mappings, on définit *la jointure* de Ω_1 et Ω_2 par: $\Omega_1 \bowtie \Omega_2 = \{\mu_1 \cup \mu_2 \mid \mu_1 \in \Omega_1, \mu_2 \in \Omega_2 \; sont \; des \; mappings \; compatibles\}$. Soit D une base de données RDF, t un patron de triplet. L'évaluation du patron de triplet t dans D notée $[[t]]_D$ est définie par: $[[t]]_D = \{\mu \mid dom(\mu) = var(t) \wedge \mu(t) \in D\}$. soit Q une requête, l'évaluation de Q sur D est définie par: $[[Q]]_D = [[t_1]]_D \bowtie \cdots \bowtie [[t_n]]_D$. Cette évaluation peut être effectuée sous différents régimes d'implication tels que définis dans la spécification $SPARQL$ (par exemple, le régime d'implication simple ou RDF). Soit μ une solution de la requête $Q = t_1 \wedge \cdots \wedge t_n$ et $aggreg$ une fonction d'agrégation (par exemple, le min), le degré de confiance de μ est défini par $tv(\mu, Q) = aggreg(tv(\mu(t_1)), \cdots, tv(\mu(t_n)))$. L'évaluation de Q sur D retournant les résultats pondérés par une confiance pour un seuil α est définie par: $[[Q]]^{\alpha}_D = \{\mu \in [[Q]]_D \mid tv(\mu) \geq \alpha\}$.

Notions de αMFS et αXSS. Soit une requête $Q = t_1 \wedge \cdots \wedge t_n$, une requête $Q' = t_i \wedge \cdots \wedge t_j$ est une *sous requête* de Q, i.e. $Q' \subseteq Q$, ssi $\{i, \cdots, j\} \subseteq \{1, \cdots, n\}$. Si $\{i, \cdots, j\} \subset \{1, \cdots, n\}$, Q' est dite une *sous requête propre* de Q ($Q' \subset Q$).

DÉFINITION 1. — *Une requête minimale échouant MFS d'une requête Q est définie comme suit: $[[MFS]]_D = \emptyset \wedge \nexists \, Q' \subset MFS \; tel \; que \; [[Q']]_D = \emptyset$. Par extension, une α requête minimale échouant (αMFS) MFS d'une requête Q pour un α donné est définie par: $[[MFS]]^{\alpha}_D = \emptyset \wedge \nexists \, Q' \subset MFS \; tel \; que \; [[Q']]^{\alpha}_D = \emptyset$. L'ensemble de toutes les αMFS de Q est noté par $mfs^{\alpha}(Q)$.*

DÉFINITION 2. — *Une requête maximale réussissant XSS d'une requête Q est définie comme suit: $[[XSS]]_D \neq \emptyset \wedge \nexists \, Q' \; tel \; que \; XSS \subset Q' \wedge [[Q']]_D \neq \emptyset$. Par extension, une α requêtes maximale réussissant (αXSS) XSS de Q pour un α donné*

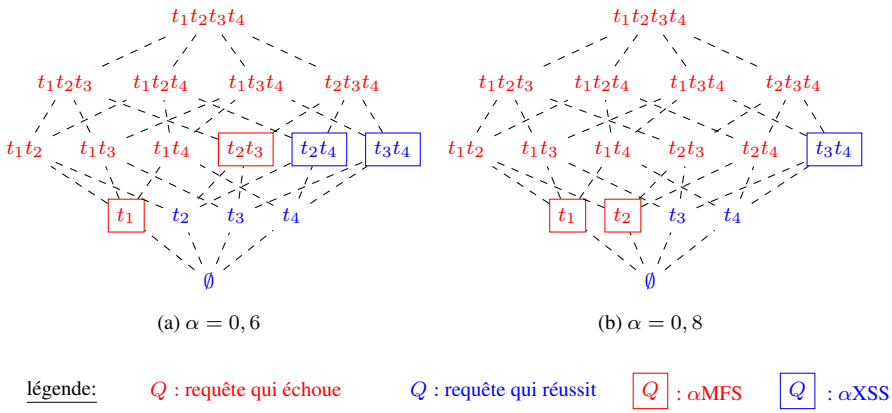


Figure 1. Treillis des sous-requêtes de Q pour différents α

est définie par: $[[XSS]]^{\alpha}_{D} \neq \emptyset \wedge \nexists\, Q'$ *tel que* $XSS \subset Q' \wedge [[Q']]^{\alpha}_{D} \neq \emptyset$. *L'ensemble de toutes les* α*XSS de* Q *pour un* α *donné est noté par* $xss^{\alpha}(Q)$.

Problématique. Dans le cadre des BC incertaines nous nous intéressons au calcul des $mfs^{\alpha_i}(Q)$ et des $xss^{\alpha_i}(Q)$ pour un ensemble de seuils $\{\alpha_1, \cdots, \alpha_n\}$.

3. Motivation

Considérons la requête suivante qui recherche les livres édités par Springer et écrits par Smith avec leurs nombres de pages. Nous désignons cette requête par $Q = t_1 \wedge t_2 \wedge t_3 \wedge t_4$ (ou $t_1t_2t_3t_4$ pour simplifier).

```
SELECT ?b ?p WHERE {
?b authors "Smith".          (t1)
?b editor "Springer" .       (t2)
?b type Book .               (t3)
?b nbPages ?p }              (t4)
```

Les αMFS et αXSS pour deux seuils de cette requête sont présentées dans la figure 1 sous forme de treillis des sous-requêtes de Q.

Supposons que l'utilisateur souhaite avoir des résultats avec un degré de confiance d'au moins 0,8. Dans notre exemple, cette requête échoue. Cependant, grâce aux 0,8 αMFS et αXSS, nous pouvons fournir les explications suivantes à l'utilisateur pour un degré de confiance de 0,8 :

1. Il n'existe aucun élément écrit par Smith.
2. Il n'existe aucun élément édité par Springer.
3. Il existe cependant des livres avec leurs nombres de pages.

Si nous calculons les αMFS et αXSS pour le degré 0,6, en plus des retours n° 1 et 3 précédents :

1. Il n'existe aucun livre édité par Springer.
2. Il existe cependant des éléments édités par Springer avec leur nombre de page.

Ces retours peuvent aider l'utilisateur à reformuler sa requête, à ajuster ses attentes ou à mieux appréhender le contenu de la base de connaissances.

4. Calcul des αMFS et αXSS pour un seuil α

Dans cette section, nous proposons dans un premier temps une adaptation de l'algorithme *Lattice-Based Approach* (LBA) proposé dans (Fokou *et al.*, 2017) pour calculer les αMFS et αXSS d'une requête pour un α donné. Notre adaptation possède la même complexité algorithmique que l'algorithme original. Notons que nous aurions aussi pu utiliser d'autres algorithmes et notamment ceux conçus pour la découverte d'ensembles fréquents maximaux comme, par exemple, *Dualize and Advance* (Gunopulos *et al.*, 2003). L'algorithme LBA présente l'avantage d'être spécifiquement optimisé pour la découverte des MFS et XSS.

4.1. L'approche αLBA

Nous présentons ici le fonctionnement de notre algorithme αLBA comme une adaptation directe de LBA dans le contexte des BC incertaines. αLBA explore le treillis des sous-requêtes d'une requête Q en suivant trois étapes.

Algorithme 1 : Découverte d'une αMFS pour une requête Q qui échoue

TrouverUneαMFS(Q, D, α)
- **entrées :** Une requête qui échoue $Q = t_1 \wedge ... \wedge t_n$;
  une base de données RDF D;
  un seuil α
- **sorties :** Une αMFS de Q notée par Q^*

1 $Q^* \leftarrow \emptyset$; $Q' \leftarrow Q$;
2 **foreach** *patron de triplet* $t_i \in Q$ **do**
3 $Q' \leftarrow Q' - t_i$;
4 **if** $[[Q' \wedge Q^*]]_D^{\alpha} \neq \emptyset$ **then**
5 $Q^* \leftarrow Q^* \wedge t_i$;
6 **return** Q^*;

1. Trouver une αMFS Q^* de Q. Suivant l'algorithme 1, αLBA supprime itérativement chaque patron de triplet t_i de Q, ce qui l'emmène à évaluer des sous-requêtes propres $Q' \wedge Q^*$ de Q. Si $Q' \wedge Q^*$ échoue pour α, alors $Q' \wedge Q^*$ contient une αMFS. Inversement, si $Q' \wedge Q^*$ réussit, alors chaque αMFS de Q contient t_i. La preuve de cette propriété repose sur le fait qu'une requête qui réussit ne peut pas contenir une requête qui échoue.

2. Calculer les α*XSS* potentielles c'est-à-dire les requêtes maximales qui ne sont pas des super-requêtes de la αMFS précédemment trouvée. L'ensemble des α*XSS potentielles* est noté $pxss(Q, Q^*)$. Cet ensemble peut être calculé comme suit:

$$pxss(Q, Q^*) = \begin{cases} \emptyset, & \text{si } |Q| = 1. \\ \{Q - t_i \mid t_i \in Q^*\}, & \text{sinon.} \end{cases}$$

Cette deuxième étape est basée sur le fait que toutes les super-requêtes de Q^* (c'est-à-dire les requêtes incluant la αMFS Q^* identifiée à l'étape précédente) renvoient un ensemble vide de réponses et peuvent donc être retirées de l'espace de recherche. Cette propriété est toujours vraie dans le contexte des BC classiques mais pour les BC incertaines. En effet, il est nécessaire qu'une requête, qui réussisse, ne puisse pas contenir de sous-requête qui échoue pour le α donné. Cette condition est détaillée dans la section 4.2

3. Tester les potentielles αXSS. Si une sous-requête trouvée lors de l'étape 2 réussit, alors il s'agit d'une αXSS. Sinon, nous appliquons les deux étapes précédentes sur cette sous-requête pour trouver une nouvelle αMFS ainsi que de nouvelles αXSS potentielles qui lui sont associées. L'algorithme 2 illustre cette étape.

Algorithme 2 : Trouver les αMFS et αXSS d'une requête Q

α**LBA(**Q**,** D**,** α**)**
- **entrées :** Une requête qui échoue $Q = t_1 \wedge ... \wedge t_n$;
  une base de données RDF D;
  un seuil α
- **sorties :** les αMFS et αXSS de Q

1 $pxss \leftarrow \{Q\}$; $mfs^\alpha(Q) \leftarrow \emptyset$; $xss^\alpha(Q) \leftarrow \emptyset$;
2 **while** $pxss \neq \emptyset$ **do**
3 $Q' \leftarrow pxss.element()$; `// choisir une` xss `potentielle`
4 **if** $[[Q']]^\alpha_D \neq \emptyset$ **then** `//` Q' `est une` α`XSS`
5 $xss^\alpha(Q) \leftarrow xss^\alpha(Q) \cup \{Q'\}$; $pxss \leftarrow pxss - \{Q'\}$;
6 **else** `//` Q' `contient au moins une` α`MFS`
7 $Q^* \leftarrow$TrouverUneαMFS(Q', D, α);
8 $mfs^\alpha(Q) \leftarrow mfs^\alpha(Q) \cup \{Q^*\}$;
`// mise à jour des` $pxss$
9 **foreach** $Q'' \in pxss$ *tel que* $Q^* \subseteq Q''$ **do**
10 $pxss \leftarrow pxss - \{Q''\}$;
11 $pxss \leftarrow pxss \cup \{Q_j \in pxss(Q'', Q^*) \mid \nexists Q_k \in pxss \cup xss^\alpha(Q)$ *tel que* $Q_j \subseteq Q_k\}$;
12 **return** $\{mfs^\alpha(Q), xss^\alpha(Q)\}$;

BC incertaine			
sujet	**prédicat**	**object**	**tv**
b_1	type	Book	0.3
b_1	nbPages	90	0.3
b_2	type	Book	0.3
b_2	nbPages	90	0.9
b_3	type	Book	0.2
b_3	nbPages	88	0.9
b_4	type	Book	0.1
b_4	nbPages	90	0.6
b_5	type	Website	0.8
b_5	nbPages	90	0.9

(a) Une base de données RDF incertaine D

```
Q : SELECT ?b WHERE {
    ?b type Book . ?b nbPages 90 }
Q' :SELECT ?b WHERE {
    ?b type Book }
```

(b) La requête Q et sa sous-requête Q'

aggreg	$[[Q']]_D^{0,4}$	$[[Q]]_D^{0,4}$
min	$\emptyset$	$\emptyset$
max	$\emptyset$	$\{b_2, b_4\}$
$\prod$	$\emptyset$	$\emptyset$
avg	$\emptyset$	$\{b_2\}$

(c) Résultats de Q et Q'

Figure 2. Illustration des différentes fonctions d'agrégation

4.2. Contraintes sur la fonction d'agrégation

Comme indiqué dans la section précédente, l'algorithme αLBA repose sur le fait qu'une requête qui réussit ne peut pas contenir une requête qui échoue. Dans le contexte des BC incertaines, selon la fonction d'agrégation ($aggreg$) choisie, cette propriété n'est pas toujours vraie. Ceci est illustré dans la figure 2, où nous présentons les résultats d'une requête Q et de sa sous-requête Q' sur une base de données RDF incertaine pour $\alpha = 0,4$. Pour les fonctions d'agrégation max et avg, Q réussit alors que sa sous-requête Q' échoue. Ainsi, l'algorithme αLBA ne peut pas être utilisé avec ces fonctions.

DÉFINITION 3. — *Soit* $aggreg : [0,1]^n \rightarrow [0,1]$ *une fonction d'agrégation,* $aggreg$ *est décroissante par rapport à l'inclusion ensembliste*[1] *si pour tout ensemble A et B* $\in [0,1]^n$, $A \subseteq B \Rightarrow aggreg(A) \geq aggreg(B)$.

Par définition, une fonction décroissante est monotone. Les fonctions *minimum* et *produit* appliquées aux valeurs $\in [0,1]$ sont des exemples de fonctions d'agrégation décroissantes.

PROPOSITION 4. — *Soit* $aggreg$ *une fonction décroissante. Si une sous-requête propre* Q' *de* Q *échoue pour un* α *donné (utilisant la fonction* $aggreg$*) alors* Q *échoue également pour* α.

PREUVE 5. — On considère $Q = t_1 \wedge \cdots \wedge t_n$ et sa sous-requête propre $Q' = t_i \wedge \cdots \wedge t_j$ ($\{i, \cdots, j\} \subset \{1, \cdots, n\}$). Supposons que Q' échoue et que Q réussisse : $[[Q']]_D^\alpha = \emptyset$ et $[[Q]]_D^\alpha \neq \emptyset$. Donc, $\exists \mu \in [[Q]]_D^\alpha$. Étant donné que $[[Q]]_D^\alpha \subseteq [[Q]]_D$ et $[[Q]]_D \subset [[Q']]_D$, nous avons $\mu_{|var(Q')} \in [[Q']]_D$ où $\mu_{|var(Q')}$ est la restriction de la fonction μ aux variables de Q'. Par définition, $tv(\mu, Q) = aggreg(tv(\mu(t_1)), \cdots, tv(\mu(t_n))) \geq \alpha$ et $tv(\mu_{|var(Q')}, Q') =$

1. Pour simplifier, cette définition est limitée aux ensembles mais pourrait être étendue aux multisets

$aggreg(tv(\mu(t_i)), \cdots, tv(\mu(t_j)))$ (en effet, $tv(\mu, Q') = tv(\mu_{|var(Q')}, Q')$). Étant donné que $aggreg$ est décroissante, $aggreg(tv(\mu(t_i)), \cdots, tv(\mu(t_j))) \geq aggreg(tv(\mu(t_1)), \cdots, tv(\mu(t_n))) \geq \alpha$. En conséquence, $tv(\mu_{|var(Q')}, Q') \geq \alpha$ et étant donné que $\mu_{|var(Q')} \in [[Q']]_D$ nous déduisons que $\mu_{|var(Q')} \in [[Q']]_D^{\alpha}$. Cela contredit l'hypothèse que Q' échoue. ■

Nous passons maintenant au problème de l'identification des αMFS et αXSS pour un ensemble de seuils α. En plus des causes d'échec basées sur les patrons de triplet, les retours sur des seuils multiples peuvent permettre à l'utilisateur de réévaluer le seuil de confiance associé à sa requête.

5. Calcul des αMFS et αXSS pour différents seuils α

Afin de trouver les αMFS et αXSS pour un ensemble de α: $\{\alpha_1, \cdots \alpha_n\}$, la solution naïve est d'exécuter l'algorithme αLBA pour chaque α_i. Cette solution de base est appelée $NLBA$. Dans cette section, nous discutons différentes améliorations à cette approche. L'idée est d'utiliser les αMFS et αXSS trouvées pour un seuil donné pour déduire celles d'un seuil supérieur (ou inférieur). Nous commençons par étudier une approche ascendante, pour laquelle les seuils sont considérés par ordre croissant.

5.1. Approche ascendante

Dans cette section, nous considérons deux seuils α_i et α_j tel que $\alpha_i < \alpha_j$. Si Q^* est une α_iMFS de Q, alors Q^* échoue également pour α_j. Cependant, cette sous-requête n'est pas nécessairement minimale pour α_j et peut donc ne pas être une α_jMFS. La proposition suivante énonce une condition à laquelle une α_iMFS est aussi une α_jMFS.

PROPOSITION 6. — *Soit α_i et α_j deux seuils tel que $\alpha_i < \alpha_j$ et Q^* une α_iMFS de Q sur un ensemble de données D. Si $|Q^*| = 1$, alors Q^* est aussi une α_jMFS de Q.*

PREUVE 7. — Si Q^* est une α_iMFS de Q sur un ensemble de données D, alors $[[Q^*]]_D^{\alpha_i} = \emptyset$. Puisque $\alpha_i < \alpha_j$, nous avons aussi $[[Q^*]]_D^{\alpha_j} = \emptyset$. Q^* est minimale ($|Q^*| = 1$) et échoue pour α_j, ainsi Q^* est une α_jMFS de Q. ■

Vérifier si une requête ne possède qu'un patron de triplet ne nécessite aucun accès à la BC. Ainsi, nous vérifions d'abord ce cas simple et ajoutons le cas échéant les α_jMFS correspondantes à l'ensemble des *αMFS découvertes*, noté $dmfs^{\alpha_j}(Q)$. Dans le cas contraire ($|Q^*| \geq 2$), prouver que Q^* est une α_jMFS nécessite de vérifier la réussite de toutes ses sous-requêtes, ce qui exige l'exécution de toutes celles-ci (soit $|Q^*|$ requêtes) sans garantie de trouver une α_jMFS. En revanche, l'algorithme TrouverUneαMFS de αLBA (algorithme 1) requiert lui aussi l'exécution de $|Q^*|$ requêtes mais garantit qu'une αMFS soit trouvée. Ainsi, notre approche privilégie TrouverUneαMFS par rapport à l'exécution des sous-requêtes de la α_iMFS, comme nous le verrons dans l'algorithme 3.

Quant aux αXSS, une α_iXSS de Q ne réussit pas nécessairement pour le seuil α_j. La proposition suivante indique que si cette α_iXSS réussit, elle est aussi une α_jXSS.

PROPOSITION 8. — *Soit α_i et α_j deux seuils tel que $\alpha_i < \alpha_j$ et Q^* une α_iXSS de Q sur un ensemble de données D. Si $[[Q^*]]_D^{\alpha_j} \neq \emptyset$, alors Q^* est une α_jXSS de Q.*

PREUVE 9. — Si Q^* est une α_iXSS de Q sur un ensemble de données D, alors toutes ses super-requêtes échouent pour α_i (autrement, elle ne serait pas maximale). Comme $\alpha_i < \alpha_j$, ses super-requêtes échouent également pour α_j. Si $[[Q^*]]_D^{\alpha_j} \neq \emptyset$, Q^* réussit et elle est maximale pour α_j. Ainsi, Q^* est une α_jXSS de Q. ■

Ainsi, la vérification qu'une α_iXSS soit aussi une α_jXSS requiert l'exécution d'une seule requête. Ceci nous permet de trouver un ensemble de α_jXSS découvertes, notés $dxss^{\alpha_j}(Q)$.

L'algorithme 3 présente notre approche complète pour trouver un ensemble de α_jMFS et de α_jXSS en se basant sur les α_iMFS et α_iXSS. Toutes les α_iMFS composées d'un seul patron de triplet (*requêtesElementaire*) sont insérées dans $dmfs^{\alpha_j}(Q)$ (ligne 1). L'algorithme itère ensuite sur les α_iMFS possédant au moins deux patrons de triplet (l'ensemble FQ) pour lesquelles il cherche une α_jMFS Q^* (ligne 6). Les sur-requêtes de Q^* sont alors retirées de l'ensemble FQ car elles ne peuvent plus être minimales. Les α_jXSS sont ensuite identifiées en exécutant chaque α_iXSS et en conservant celles qui réussissent pour le seuil α_j (lignes 10-11).

Algorithme 3 : Découvrir des α_jMFS et α_jXSS pour l'approche ascendante

DécouvrirαMFSXSS($mfs^{\alpha_i}(Q)$**,** $xss^{\alpha_i}(Q)$ D**,** α_j**)**
- **entrées :** Les α_iMFS $mfs^{\alpha_i}(Q)$ d'une requête Q pour un seuil α_i;
  Les α_iXSS $xss^{\alpha_i}(Q)$ d'une requête Q pour un seuil α_i;
  une base de données RDF D;
  un seuil $\alpha_j > \alpha_i$
- **sorties :** Un ensemble de α_jMFS de Q noté $dmfs^{\alpha_j}(Q)$;
  Un ensemble de α_jXSS de Q noté $dxss^{\alpha_j}(Q)$;

1 *requêtesElementaire* $\leftarrow \{Q_a \in mfs^{\alpha_i}(Q) \mid |Q_a| = 1\}$;
2 $dmfs^{\alpha_j}(Q) \leftarrow$ *requêtesElementaire*;
3 $FQ \leftarrow mfs^{\alpha_i}(Q) -$ *requêtesElementaire*;
4 **while** $FQ \neq \emptyset$ **do**
5 $\quad Q' \leftarrow fQ.dequeue()$;
6 $\quad Q^* \leftarrow TrouverUne\alpha MFS(Q', D, \alpha_j)$;
7 $\quad dmfs^{\alpha_j}(Q) \leftarrow dmfs^{\alpha_j}(Q) \cup \{Q^*\}$;
8 $\quad$ **foreach** $Q'' \in FQ$ *tel que* $Q^* \subseteq Q''$ **do**
9 $\quad\quad FQ \leftarrow FQ - \{Q''\}$;
10 **foreach** $Q^* \in xss^{\alpha_i}(Q)$ *tel que* $[[Q^*]]_D^{\alpha_j} \neq \emptyset$ **do**
11 $\quad dxss^{\alpha_j}(Q) \leftarrow dxss^{\alpha_j}(Q) \cup \{Q^*\}$;
12 **return** $\{dmfs^{\alpha_j}(Q), dxss^{\alpha_j}(Q)\}$;

Après avoir découvert les α_jMFS et α_jXSS, une version optimisée de αLBA est exécutée en prenant en entrée les α_jMFS et α_jXSS découvertes (voir l'algorithme 4). Cet algorithme calcule les α_jXSS potentielles ($pxss$) qui ne contiennent aucune α_jMFS (lignes 2-6), supprime de cet ensemble les α_jXSS (ligne 7), puis, itère sur l'ensemble $pxss$ comme pour la version originale de αLBA (voir l'algorithme 2).

Algorithme 4 : Version améliorée de αLBA

Optimized-αLBA(Q**,** D**,** α**,** $dmfs^{\alpha}(Q)$**,** $dxss^{\alpha}(Q)$**)**

entrées : Une requête qui échoue Q;
une base de données RDF D;
un seuil α
un ensemble de αMFS de Q noté $dmfs^{\alpha}(Q)$;
un ensemble de αXSS de Q noté $dxss^{\alpha}(Q)$;

sorties : Les αMFS et αXSS de Q

1 $mfs^{\alpha}(Q) \leftarrow dmfs^{\alpha}(Q)$; $xss^{\alpha}(Q) \leftarrow dxss^{\alpha}(Q)$;

2 $Q^* \leftarrow dmfs^{\alpha}(Q).dequeue()$; $pxss \leftarrow pxss(Q, Q^*)$;

3 **foreach** $Q^* \in dmfs^{\alpha}(Q)$ **do**

4 **foreach** $Q' \in pxss$ *tel que* $Q^* \subseteq Q'$ **do**

5 $pxss \leftarrow pxss - \{Q'\}$;

6 $pxss \leftarrow pxss \cup \{Q_i \in pxss(Q', Q^*) \mid \nexists Q_j \in pxss \cup xss^{\alpha}(Q) : Q_i \subseteq Q_j\}$;

7 $pxss \leftarrow pxss - dxss^{\alpha}(Q)$;

8 **while** $pxss \neq \emptyset$ **do**

`// idem que les lignes 3-11 de αLBA`

9 **return** $\{mfs^{\alpha}(Q), xss^{\alpha}(Q)\}$;

Pour illustrer l'approche ascendante, nous considérons à nouveau l'exemple de la figure 1. À partir des 0,6 αMFS et αXSS, nous montrons comment l'algorithme ascendant calcule les 0,8 αMFS et αXSS. Comme t_1 est une 0,6 αMFS et ne contient qu'un seul patron de triplet, cette dernière est une 0,8 αMFS (proposition 6). la seconde 0,6 αMFS est t_2t_3. Comme cette requête échoue nécessairement pour 0,8, nous utilisons l'algorithme $TrouverUne\alpha MFS$ pour identifier la 0,8 αMFS t_2. Compte tenu des 0,6 αXSS, seule t_3t_4 réussit pour 0,8. Ainsi, t_3t_4 est une 0,8 αXSS (proposition 8). Les αMFS et αXSS découvertes données en entrée de l'algorithme 4 sont respectivement: $dmfs^{\alpha}(Q) = \{t_1, t_2\}$ et $dxss^{\alpha}(Q) = \{t_3t_4\}$. À partir de ces ensembles l'algorithme 4 détermine qu'il n'éxiste pas de αXSS potentielles (lignes 1-7) et ainsi, que toutes les 0,8 αMFS et αXSS ont été trouvées.

5.2. Approche descendante

L'approche ascendante nous permet de découvrir quelques αMFS et αXSS (et ainsi, améliorer l'exécution de αLBA pour plusieurs seuils α), pour des valeurs de seuils croissantes. Nous proposons ici une approche descendante qui calcule les αMFS

et αXSS en utilisant des valeurs de seuil décroissantes. Grâce à la relation de dualité entre les αMFS et αXSS, les propriétés et algorithmes utilisés pour cette approche sont similaires à ceux de l'approche ascendante. Soit α_i et α_j deux seuils tel que $\alpha_i > \alpha_j$,

– les α_iMFS qui échouent pour α_j sont des α_jMFS;

– les α_iXSS d'une taille $|Q| - 1$ sont des α_jXSS;

– les α_iXSS d'une taille $< |Q| - 1$ réussissent aussi pour α_j et contiennent donc une α_jXSS.

Une fois un ensemble de α_jMFS et de α_jXSS découvert, l'algorithme Optimized-αLBA (algorithme 4) est exécuté pour trouver les α_jMFS et α_jXSS restantes.

6. Évaluation expérimentale

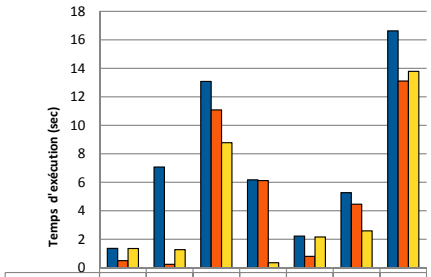
Dans cette section, nous étudions les performances de nos deux approches en les comparant avec l'approche de base $NLBA$ (exécution de αLBA pour chacun des N seuils).

Environnement expérimental. Nous avons implémenté nos algorithmes avec JAVA 1.8 64 bits. Les algorithmes prennent en entrée une requête Q qui échoue et un ensemble de seuils, et renvoient les ensembles de αMFS et αXSS de Q pour chaque seuil. Dans l'implémentation actuelle, les requêtes sont exécutées avec Jena TDB. Nous avons choisi Jena TDB car ce Quadstore nous permet de stocker le degré de confiance associé à chaque triplet. Jena TDB fournit un filtre de bas niveau permettant de récupérer les résultats satisfaisant le seuil fourni. Notre implémentation est disponible sur https://forge.lias-lab.fr/projects/qars4ukb avec un tutoriel pour reproduire nos expérimentations.

Nos expérimentations ont été menées sur un système Ubuntu Server 16.04 LTS avec un processeur Intel XEON E5-2630 v3 @2.4Ghz et 16GB de RAM. Arbitrairement, nous utilisons la fonction d'agrégation min. Les temps d'exécution sont une moyenne de cinq exécutions consécutives. Pour éviter un effet de démarrage à froid, une exécution préliminaire est effectuée mais non considérée.

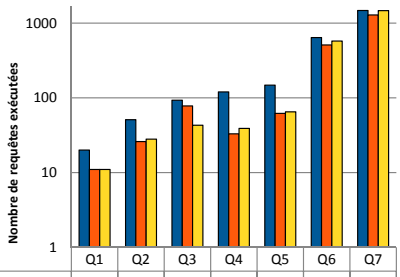
Données et requêtes. Nous avons utilisé un jeu de données de 20M triplets générés par le Benchmark WatDiv (Aluç *et al.*, 2014). Le degré de confiance de chaque triplet RDF a été généré aléatoirement avec une distribution uniforme sur [0 , 1]. Nous considérons 7 requêtes qui échouent (voir Tableau 1 en annexe). Ces requêtes contiennent entre 1 et 15 patrons de triplet et couvrent les principaux types de requêtes: étoile (jointure *sujet-sujet* entre les patrons de triplet), chaîne (jointure *objet-objet*) et composite (d'autres jointures).

Expérimentation. Notre expérimentation évalue nos algorithmes pour les seuils $\{0,2; 0,4; 0,6; 0,8\}$ et un jeu de données de 20M. La figure 3a montre le temps d'exécution de chaque algorithme pour chaque requête. La figure 3b donne le nombre de requêtes exécutées.



NLBA	1,36	7,07	13,08	6,17	2,22	5,27	16,63
Ascendant	0,5	0,24	11,08	6,12	0,8	4,46	13,11
Descendant	1,35	1,27	8,78	0,35	2,16	2,59	13,79

(a) Temps d'exécution (20M triplets)



	Q1	Q2	Q3	Q4	Q5	Q6	Q7
NLBA	20	51	93	120	148	641	1475
Ascendant	11	26	78	33	62	510	1286
Descendant	11	28	43	39	65	576	1470

(b) # Requêtes exécutées

Figure 3. Résultats expérimentaux avec Jena TDB

Cette expérimentation montre les gains apportés par nos algorithmes par rapport à la méthode de base NLBA : nos algorithmes exécutent moins de requêtes pour trouver les αMFS et αXSS, soit en moyenne respectivement 39% et 40% moins de requêtes. Par conséquent, ces algorithmes ont des temps d'exécutions inférieurs, réduisant ainsi le temps d'exécution total de 30% et 42% respectivement. Par exemple, NLBA a besoin de 7 secondes pour trouver les αMFS et αXSS de la requête Q2, alors que nos algorithmes effectuent ce calcul en environ 1 seconde. La différence entre les temps d'exécutions dépend fortement des requêtes que nos algorithmes évitent d'exécuter. Par exemple, nos algorithmes exécutent moins de 40 requêtes pour Q4 tandis que NLBA en exécute 120, pourtant, l'approche ascendante améliore le temps d'exécution de moins de 1%. Pour l'algorithme descendant, le gain de performance est significatif (94%). En analysant les requêtes exécutées, nous constatons que l'algorithme ascendant évite d'exécuter des requêtes qui ont des temps d'exécution courts, et qu'il exécute tout de même les requêtes les plus coûteuses. Ainsi, le temps d'exécution global reste quasiment inchangé. Cette expérimentation montre également qu'aucun de nos algorithmes ne fournit le meilleur résultat pour toutes les requêtes : l'approche ascendante est meilleure pour Q1, Q2 et Q5 tandis que l'approche descendante est meilleure pour Q3, Q4 et Q6. Malgré l'exécution de peu de requêtes, l'algorithme ascendant possède un temps d'exécution total supérieur à l'algorithme descendant. Ceci est dû au fait que ces algorithmes exécutent différentes requêtes avec des temps d'exécution différents. En particulier, l'algorithme descendant commence par les seuils les plus élevés. Les requêtes exécutées tendent à être sélectives étant donné que le critère de confiance est restrictif, et par conséquent, elles ont des temps d'exécution courts. Une fois les αMFS et αXSS trouvées pour les seuils les plus élevés, l'algorithme descendant évite ainsi l'exécution de requêtes avec un seuil inférieur, susceptibles d'être plus coûteuses. Comme l'algorithme ascendant suit l'approche duale, il tend à exécuter des requêtes peu sélectives avec des coûts plus importants.

7. Travaux Connexes

Dans le contexte des BC, le problème des réponses vides a été abordé par plusieurs approches complémentaires telles que la complétion de la BC en utilisant des règles de déduction logiques (Galárraga *et al.*, 2015), la vérification des données lors de la formulation de la requête (Campinas, 2014), un schéma relationnel émerge à partir des données de la BC pour aider les utilisateurs à formuler des requêtes (Pham *et al.*, 2015) ou à relaxer la requête pour retourner des réponses alternatives (Hurtado *et al.*, 2009 ; Huang *et al.*, 2012 ; Fokou *et al.*, 2014 ; Calí *et al.*, 2014 ; Hogan *et al.*, 2012 ; Elbassuoni *et al.*, 2011 ; Dolog *et al.*, 2009). Dans cette section, nous résumons les principales contributions sur la relaxation des requêtes RDF.

Il existe plusieurs travaux qui ont proposé des opérateurs de relaxation dans le contexte RDF. Ces opérateurs sont principalement basés sur la sémantique RDFS (par exemple, la généralisation des patrons de triplet utilisant des hiérarchies de classes et de propriétés) (Hurtado *et al.*, 2009 ; Huang *et al.*, 2012 ; Fokou *et al.*, 2014 ; Calí *et al.*, 2014), les mesures de similarité (Hogan *et al.*, 2012 ; Elbassuoni *et al.*, 2011) et les préférences utilisateur (Dolog *et al.*, 2009). Ces opérateurs génèrent un ensemble de requêtes relaxées, ordonnées par similarité avec la requête d'origine et exécutées dans cet ordre (Hurtado *et al.*, 2009 ; Huang *et al.*, 2012 ; Reddy, Kumar, 2013). Les opérateurs de relaxation peuvent être directement utilisés par l'utilisateur dans sa requête (Fokou *et al.*, 2014 ; Calí *et al.*, 2014) ou utilisés conjointement avec des règles de réécriture de requêtes pour effectuer la relaxation (Dolog *et al.*, 2009). Avec ces approches, les causes d'échec de la requête sont inconnues, ce qui peut conduire à exécuter des requêtes relaxées inutiles. Fokou et al. (Fokou *et al.*, 2017 ; 2016) ont abordé ce problème en définissant d'abord les approches LBA et MBA pour calculer les MFS et XSS de la requête (Fokou *et al.*, 2017) et en proposant des stratégies de relaxation basées sur les MFS qui identifient les requêtes relaxées échouant nécessairement (Fokou *et al.*, 2016). Notre approche est basée sur l'algorithme LBA proposé dans ce travail, que nous avons étendu en identifiant la condition pour laquelle cet algorithme peut être utilisé dans le contexte des BC incertaines. Notre travail est parmi les travaux pionniers visant à explorer le problème de la relaxation de requêtes dans le contexte des BC incertaines. Pour autant que nous sachions, le seul autre travail déjà réalisé dans ce contexte est (Reddy, Kumar, 2013). toutefois, ce travail utilise uniquement la valeur de confiance pour ordonner les résultats par leur fiabilité. Ils ne considèrent pas, comme nous le faisons dans cet article, les requêtes qui ne renvoient aucun résultat satisfaisant le degré de confiance fourni.

Enfin, nous notons que le problème original de découverte des MFS et des XSS d'une requête SPARQL est analogue à la découverte d'ensembles fréquents maximaux (Mannila, Toivonen, 1997 ; Gunopulos *et al.*, 2003). En effet, les deux problèmes reviennent à chercher des bordures positives et négatives d'une propriété monotone dans l'espace des solutions représenté par un treillis. Cependant, nous nous intéressons principalement dans cet article à un nouveau problème de découverte des MFS et XSS pour plusieurs seuils. Ce problème n'est pas équivalent à la découverte d'ensembles fréquents maximaux car plusieurs treillis doivent être considérés. D'un point

de vue théorique, celui-ci s'apparente à la recherche d'ensembles fréquents maximaux pour plusieurs seuils de fréquence, qui est à notre connaissance peu étudié dans la littérature.

8. Conclusion

Dans cet article, nous nous somme intéressés au problème des réponses vides dans le contexte des BC incertaines où une requête échoue si elle ne renvoie aucun résultat ou si elle renvoie un résultat qui ne satisfait pas le degré de confiance α attendu. Pour fournir à l'utilisateur un retour pertinent, nous avons proposé de calculer les αMFS et αXSS de la requête, car elles donnent un aperçu clair des causes d'échec et un ensemble de requêtes alternatives retournant des résultats utiles.

Nous avons d'abord défini la condition pour laquelle l'algorithme de nos travaux précédents, appelé αLBA, peut être directement adapté au contexte des BC incertaines. Dans ce cas, l'utilisateur doit définir un degré de confiance attendu. Cependant, l'utilisateur peut vouloir savoir ce qui se passe s'il assouplit cette condition sur la confiance. Ainsi, nous avons étudié le problème du calcul des αMFS et αXSS pour plusieurs seuils. La méthode de base, appelée NLBA, consiste à exécuter αLBA pour chaque seuil. Cependant, nous avons observé et prouvé que les αMFS et αXSS pour un seuil donné peuvent être réutilisées pour trouver celles d'un seuil inférieur (ou supérieur). Ainsi, nous avons défini deux approches alternatives de NLBA, appelées ascendante et descendante, qui considèrent des seuils α dans l'ordre croissant ou décroissant. Nous avons effectué une mise en œuvre complète de ces algorithmes et montré expérimentalement sur différents jeux de données du benchmark WatDiv que nos approches sont plus performantes que la méthode de base.

À court terme, nous envisageons d'illustrer l'intérêt de nos approches en l'appliquant sur un contexte réel comme, par exemple, avec des requêtes exécutées sur la base de connaissances YAGO. Dans nos expérimentations, nous avons aussi observé qu'aucun de nos algorithmes n'offre les meilleures performances pour toutes les requêtes. Comme autre perspective, nous envisageons d'étudier les conditions dans lesquelles un algorithme fournit les meilleurs résultats. Notre idée est d'utiliser les statistiques des BC et le modèle de coût du système de gestion des triplets pour trouver l'algorithme susceptible de proposer les meilleures performances. Enfin, à plus long terme, nous pensons qu'il serait intéressant d'adapter nos approches au contexte des données massives.

Bibliographie

Aluç G., Hartig O., Özsu M. T., Daudjee K. (2014). Diversified Stress Testing of RDF Data Management Systems. In *Iswc'14*, p. 197–212.

Calí A., Frosini R., Poulovassilis A., Wood P. (2014). Flexible Querying for SPARQL. In *Odbase'14*, p. 473-490.

Campinas S. (2014). Live SPARQL Auto-Completion. In *Iswc'14*, p. 477–480.

DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia. (2015). *Semantic Web*, vol. 6, n° 2, p. 167–195.

Dolog P., Stuckenschmidt H., Wache H., Diederich J. (2009). Relaxing RDF queries based on user and domain preferences. *Journal of Intelligent Information Systems (JIIS)*, vol. 33, n° 3, p. 239–260.

Elbassuoni S., Ramanath M., Weikum G. (2011). Query Relaxation for Entity-Relationship Search. In *ESWC'11*, p. 62–76.

Fokou G., Jean S., Hadjali A. (2014). Endowing Semantic Query Languages with Advanced Relaxation Capabilities. In *ISMIS'14*, p. 512–517.

Fokou G., Jean S., HadjAli A., Baron M. (2016). RDF Query Relaxation Strategies Based on Failure Causes. In *Eswc'16*, p. 439–454.

Fokou G., Jean S., Hadjali A., Baron M. (2017). Handling Failing RDF Queries: From Diagnosis to Relaxation. *Knowledge and Information Systems (KAIS)*, vol. 50, n° 1.

Galárraga L., Teflioudi C., Hose K., Suchanek F. M. (2015). Fast Rule Mining in Ontological Knowledge Bases with AMIE+. *VLDB Journal*, vol. 24, n° 6, p. 707–730.

Gunopulos D., Khardon R., Mannila H., Saluja S., Toivonen H., Sharm R. S. (2003). Discovering All Most Specific Sentences. *ACM Trans. on Database Systems*, vol. 28, n° 2, p. 140–174.

Hartig O. (2009). Querying Trust in RDF Data with tSPARQL. In *ESWC 2009*.

Hogan A., Mellotte M., Powell G., Stampouli D. (2012). Towards Fuzzy Query-relaxation for RDF. In *ESWC'12*, p. 687–702.

Huang H., Liu C., Zhou X. (2012). Approximating query answering on RDF databases. *Journal of the World Wide Web*, vol. 15, n° 1, p. 89–114.

Hurtado C. A., Poulovassilis A., Wood P. T. (2009). Ranking Approximate Answers to Semantic Web Queries. In *ESWC'09*, p. 263-277.

Knowledge Vault: A Web-scale Approach to Probabilistic Knowledge Fusion. (2014). In *Kdd'14*, p. 601–610. Dong, Xin and Gabrilovich, Evgeniy and Heitz, Geremy and Horn, Wilko and Lao, Ni and Murphy, Kevin and Strohmann, Thomas and Sun, Shaohua and Zhang, Wei.

Mannila H., Toivonen H. (1997). Levelwise Search and Borders of Theories in Knowledge Discovery. *Data Mining and Knowledge Discovery*, vol. 1, n° 3, p. 241–258.

Pérez J., Arenas M., Gutierrez C. (2009). Semantics and Complexity of SPARQL. *ACM Transaction on Database Systems (TODS)*, vol. 34, n° 3, p. 16:1–16:45.

Pham M., Passing L., Erling O., Boncz P. A. (2015). Deriving an Emergent Relational Schema from RDF Data. In *Www'15*, p. 864–874.

Reddy K. B., Kumar P. S. (2013). Efficient Trust-Based Approximate SPARQL Querying of the Web of Linked Data. In *Uncertainty Reasoning for the Semantic Web II*, p. 315–330.

Saleem M., Ali M. I., Hogan A., Mehmood Q., Ngomo A. N. (2015). LSQ: The Linked SPARQL Queries Dataset. In *Iswc'15*, p. 261–269.

Tomaszuk D., Pak K., Rybiński H. (2013). Trust in RDF graphs. In *Adbis'13*.

Annexe

Tableau 1. Requêtes utilisées pour l'évaluation expérimentale (en format simplifié)

Q1 (3TP*)	`SELECT * WHERE { ?p friendOf ?f . ?f likes ?p . ?p type ProductCategory }`
Q2 (6TP)	`SELECT * WHERE { User666524 likes ?v0 . ?v0 hasGenre ?v1 . ?v1 tag Topic129 . ?v0 friendOf ?v2 . ?v2 Location ?v3 . ?v3 parentCountry Country17 }`
Q3 (7TP)	`SELECT * WHERE { ?v0 follows ?v1 . ?v1 follows ?v0 . ?v1 subscribes ?v2 . ?v0 subscribes ?v2 . ?v1 likes Product16770 . ?v0 nationality Country20 . ?v0 makesPurchase ?v3 }`
Q4 (8TP)	`SELECT * WHERE { ?v0 type User . ?v0 familyName 'Smith' . ?v0 subscribes Website362909 . ?v0 follows ?v1 . ?v0 friendOf ?v2 . ?v0 likes ?v3 . ?v0 userId ?v4 . ?v0 makesPurchase ?v5 . ?v0 Location ?v6 . ?v0 nationality ?v7 . ?v0 userId ?v8 }`
Q5 (10TP)	`SELECT * WHERE { ?p likes ?x . ?x likes ?p . ?p hasGenre SubGenre92 . ?x subscribe ?w1 . ?w1 language Language21 . Website121 hits ?h . ?x homepage Website120 . ?x familyName 'Smith' . ?x friendOf ?x2 . ?x2 email 'xxx@xxx.com' }`
Q6 (12TP)	`SELECT * WHERE { ?v0 eligibleRegion Country05 . ?v0 includes ?v1 . Retailer1257 offers ?v0 . ?v0 price '90' . ?v0 serialNumber ?v4 . ?v0 validFrom ?v5 . ?v0 validThrough ?v6 . ?v0 eligibleQuantity ?v8 . ?v0 priceValidUntil ?v11 . ?v1 tag ?v7 . ?v1 keywords ?v10 . ?v12 purchaseFor ?v1 }`
Q7 (15TP)	`SELECT * WHERE { ?v0 type ProductCategory7 . ?v0 tag Topic245 . ?v0 hasReview ?v4 . ?v0 contentSize ?v9 . ?v0 description ?v10 . ?v0 keywords ?v11 . ?v12 purchaseFor ?v0 . ?v2 tag ?v1 . ?v4 rating ?v5 . ?v4 reviewer ?v6 . ?v4 text ?v7 . ?v4 title ?v8 . ?v6 familyName ?v13 . ?v6 birthDate ?v14 . ?v0 gender ?v15 }`

* nombre de patrons de triplet

Graphes enrichis par des Cubes (GreC) : une approche innovante pour l'OLAP sur des réseaux d'information

Cécile Favre[1], **Wararat Jakawat**[2], **Sabine Loudcher**[3]

1. *Université de Lyon, Université Lyon 2, ERIC EA 3083, France*
   cecile.favre@univ-lyon2.fr
2. *Computer Science Department, Prince of Songkhla University, Thailand*
   wararat.j@psu.ac.th
3. *Université de Lyon, Université Lyon 2, ERIC EA 3083, France*
   sabine.loudcher@univ-lyon2.fr

ABSTRACT. *In order to make the online analysis of information networks, several works combine OLAP and graphs, combination known as* Graph OLAP. *To complete these works which often offer to analyze cubes of graphs, in a different and complementary way, we propose a new approach called* GreC *(*Graphs enriched by Cubes*). Instead of building cubes of graphs, our proposal is to enrich the graphs with cubes, graphs where the nodes or edges of the network are described by cubes. This allows interesting analyses for the user who can navigate within a graph enriched by cubes according to different levels of analysis, with dedicated operators. In this article we recall the general framework of* GreC *and show how the classic concepts of OLAP must be revisited and expanded. Then we will focus on metadata in the model that ensures the genericity of the approach, on the implementation of a prototype and on elements of performances.*

RÉSUMÉ. *Afin de pouvoir faire de l'analyse en ligne de réseaux d'information, plusieurs travaux proposent de combiner l'OLAP et les graphes, combinaison connue sous le nom de* Graph OLAP. *Pour compléter ces travaux dont le principe fondamental est d'analyser des cubes de graphes, nous proposons une approche innovante appelée* GreC *(*Graphes enrichis par des Cubes*). Plutôt que de construire des cubes de graphes, notre proposition consiste à enrichir les graphes avec des cubes de données qui viennent décrire les nœuds et/ou les arêtes du réseau selon les besoins. Cela permet des analyses intéressantes pour l'utilisateur qui peut naviguer au sein d'un graphe enrichi de cubes selon différents niveaux d'analyse, avec des opérateurs dédiés. Dans cet article nous rappelons le cadre général de l'approche* GreC *et montrons comment les concepts classiques de l'OLAP doivent être revisités et étendus. Puis nous nous focalisons sur les métadonnées qui permettent d'assurer la généricité de l'approche, sur l'implémentation d'un prototype et donnons quelques éléments relatifs aux performances.*

KEYWORDS: *Réseau d'informations ; cube ; OLAP ; graph OLAP ; opération informationnelle ; opération topologique.*

MOTS-CLÉS : *Information network ; cube ; OLAP ; graph OLAP ; informational operation ; topological operation.*

1. Introduction

Historiquement, l'analyse OLAP (*Online Analytical Processing*) a été développée et utilisée dans un contexte de données assez classiques, souvent structurées dans des bases de données. L'émergence de nouveaux types de données à considérer, comme par exemple le texte, les images, les réseaux d'information, a soulevé de nouveaux défis à relever pour permettre une extension de cette technologie, entre autres en revisitant les concepts, en recherchant comment transposer ce qui existait à de nouveaux types de données, en développant de nouvelles approches prenant en compte ces nouveaux types de données, et ce pour tirer parti de la richesse de leurs spécificités. Dans le paysage des données complexes, les réseaux d'information constituent un type de données particulièrement riche compte-tenu non seulement de la multiplicité des données, mais aussi de leurs liens. La modélisation sous forme de graphes avec des nœuds et des arêtes peut prendre différentes formes selon les besoins de représentation : graphe valué ou non pour la pondération des arcs, graphe homogène (un seul type de nœud) ou hétérogène, etc.

Pour illustrer les réseaux d'information, considérons les données bibliographiques qui se prêtent particulièrement bien à la représentation sous forme de graphes. Ces données ont d'ailleurs fait l'objet des premières approches qui ont tenté de combiner les graphes et l'approche OLAP. Il apparaît qu'une des caractéristiques importantes des données bibliographiques réside dans le fait que, de par leur nature, elles sont liées entre elles et peuvent donner lieu à une représentation sous forme de graphe. Par exemple, le fait que deux auteurs aient publié ensemble induit le fait que sur un graphe d'auteurs, si nous nous intéressons à la co-publication, l'arête reliant ces deux auteurs pourra être valuée par le nombre de papiers que les personnes ont écrits ensemble. Par exemple dans la Figure 1, J. Han et Y.Sun ont collaboré au travers de 5 publications.

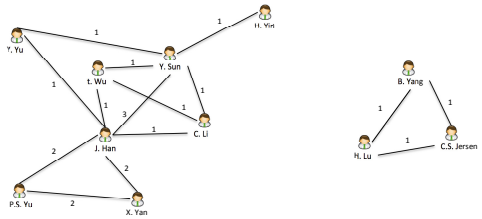


Figure 1. Graphe d'auteurs représentant les co-publications à un instant t.

Néanmoins, dans cet exemple, on peut constater que le pouvoir informatif de ce graphe reste assez faible. En effet, cette représentation ne prend pas en compte la dynamique des données (c'est une photo à un instant t de l'état des co-publications) ; par ailleurs, cette représentation ne permet pas de rendre compte de différentes informations caractérisant les publications dénombrées, telles que l'année, le lieu de publication, la thématique, etc.

Une autre alternative de visualisation pour rendre compte de ces informations correspond à ce qui est proposé par l'analyse OLAP avec une représentation multidimensionnelle sous forme de cube. Par exemple, dans la Figure 2, il est possible d'analyser le *fait* (objet d'analyse) "production scientifique", au travers d'une *mesure* (indicateur) qui est le "nombre de publications", en fonction de différentes *dimensions* (axes d'analyse) qui sont ici au nombre de trois : "auteur", "temps" et "lieu" (venue). Ces dimensions peuvent être organisées sous forme de *hiérarchies*, organisées en différents niveaux de granularité. Par exemple, la dimension "lieu" a une hiérarchie en deux niveaux : un niveau avec le nom du lieu de publication et un niveau "domaine". La présence d'une dimension hiérarchisée est un élément important de la navigation dans les données. En effet, l'OLAP traditionnel propose différentes opérations de navigation dans les données. Parmi les plus utilisées, il y a les opérations qui permettent de naviguer à travers les niveaux de détail des données selon les dimensions hiérarchisées, avec un processus d'agrégation : le *Roll Up* (forage ascendant) qui permet d'obtenir les données à un niveau agrégé et le *Drill Down* (forage descendant) qui fait l'inverse. Il est à noter que dans cette représentation multidimensionnelle qui comporte davantage d'informations, le fait que les auteurs soient en lien au travers de ces publications (co-publications) n'apparaît pas du tout.

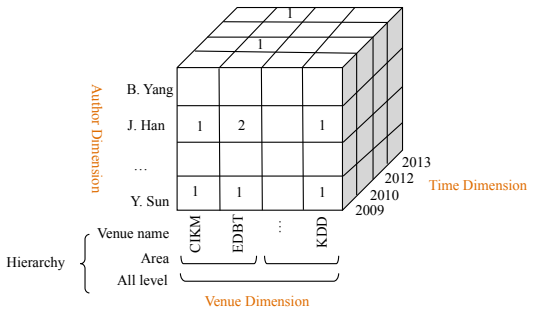


Figure 2. Structure d'un cube de données dans le contexte OLAP pour des données bibliographiques.

Ainsi, afin de tirer parti de ces deux visualisations (graphe et cube), un nouveau champ de recherche est apparu : *Graph OLAP* (Chen *et al.*, 2008). Le *Graph OLAP* a fait l'objet de plusieurs publications proposant des améliorations et des extensions (Jin *et al.*, 2010; Qu *et al.*, 2011; Zhao *et al.*, 2011). L'idée sur laquelle repose initialement le *Graph OLAP* consiste à construire un cube de graphes dans lequel il est

possible de naviguer, grâce à différents opérateurs OLAP qui ont été redéfinis pour prendre en compte ce nouveau cadre d'analyse. Dans ces approches de *Graph OLAP*, il s'agit de considérer des cubes définis selon des dimensions dites informationnelles, et les mesures contenues dans les cellules correspondent, non pas à des indicateurs numériques comme traditionnellement, mais à des graphes ou plus exactement à des sous-graphes. Par exemple, dans la Figure 3, par rapport aux données considérées ici, les dimensions informationnelles sont le temps, le lieu et les mots-clés. Ici le cube est présenté sous forme de "tranche", en considérant tous les mots-clés simultanément. Dans la cellule qui est définie par les valeurs "EDBT" pour le lieu et "2013" pour le temps, le réseau est composé des co-auteurs B. Yang et C.S. Jensen qui ont co-publié un papier (dans cette conférence, cette année-là), en considérant le nombre de papiers co-publiés qui value les arêtes du graphe.

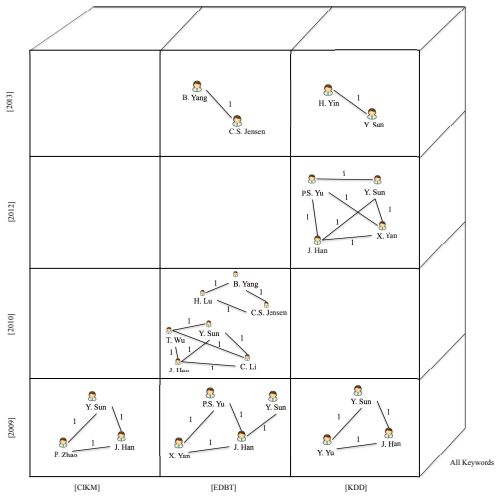


Figure 3. Cube de graphes sur des données bibliographiques pour analyser les liens de co-publication.

Dans les approches initiales de *Graph OLAP*, au niveau de la modélisation, deux types de dimensions ont été définis : les dimensions informationnelles et les dimensions topologiques. Les dimensions informationnelles vont donc conditionner les manipulations du cube. Ainsi, lors d'opérations informationnelles sur le cube, les graphes à l'intérieur des cellules vont être recalculés. Les dimensions topologiques, quant à elles, se rapportent à la modélisation des réseaux eux-mêmes dans les cellules. Les opérations topologiques sont caractérisées par un changement du type de nœuds dans les graphes. Par exemple, à partir d'un réseau d'auteurs présent dans une cellule, nous passons à un réseau d'institutions, si nous effectuons un *Roll Up* topologique selon la dimension auteur, dont la hiérarchie comprend un niveau institution.

Initialement, la combinaison de l'OLAP et des graphes s'est donc faite au travers de plusieurs approches basées sur des cubes de graphes avec dans leurs cellules, un

graphe comme mesure (Tian *et al.*, 2008; Morfonios, Koutrika, 2008; Beheshti *et al.*, 2012; Yin *et al.*, 2012). Ces approches permettent de visualiser des "instantanés" de graphes en fonction des dimensions d'analyse choisies (c'est à dire l'état d'un graphe à un instant donné). Différents opérateurs ont été proposés pour naviguer dans le cube de graphes : des opérations informationnelles ou topologiques, selon si les opérations s'appliquent selon les dimensions du cube ou les dimensions des graphes. Dans (Loudcher *et al.*, 2015), nous proposons un état de l'art et une étude comparative de ces différentes approches. Cependant, dans cette combinaison de l'OLAP et des graphes basée sur des cubes de graphes, la visualisation plus globale du graphe est perdue, alors même que celle-ci est intéressante d'un point de vue analytique. Parallèlement, la dynamique des données est importante pour l'analyse du graphe, et ceci n'est pas toujours bien perceptible dans la visualisation des parties de graphe. En effet, malgré la présence d'une dimension temporelle, en croisant celle-ci avec une ou plusieurs autres dimensions, il est difficile de se rendre compte de la dynamique même d'un graphe (évolution des arêtes ou des nœuds).

Par conséquent, nous proposons une nouvelle façon de considérer la combinaison de l'OLAP et des graphes en construisant un graphe qui réponde aux besoins de l'utilisateur avec l'enrichissement par des cubes de données pour valuer les nœuds et/ou les arêtes selon les besoins d'analyse. De plus, la présence d'une dimension temporelle dans les cubes qui valuent les nœuds et/ou les arêtes va notamment permettre de rendre compte de la dynamique du graphe. Par ailleurs, pour enrichir l'analyse, notre attention s'est focalisée sur deux apports : d'une part les types de mesures possibles ; d'autre part les opérateurs de navigation proposés. Notre approche s'intitule *GreC* pour Graphes enrichis par des Cubes.

Le cadre général de l'approche *GreC* a déjà donné lieu à des publications (Jakawat *et al.*, 2016a; 2016b). Dans le présent papier nous nous attachons à l'opérationnalité de *GreC* sur l'ensemble du processus mais aussi à sa généricité au-delà du contexte des données bibliographiques qui est le contexte initial de développement de notre approche. En effet, cette approche pourrait par exemple être appliquée dans le cadre de l'analyse de messages Twitter en se focalisant sur le graphe des *followers* enrichi par des cubes informant de l'activité propre en nombre de tweets selon le temps, la thématique, etc. ; des cubes sur les arêtes se focalisant sur les mentions entre deux *Twittos* renseignant sur la production de tweets selon les mêmes axes se restreignant aux mentions des comptes dans les tweets produits par exemple. Dans ce papier, nous gardons comme fil conducteur les données bibliographiques, tout en discutant les points permettant la généricité de l'approche. Dans cette optique, nous présentons, dans ce papier, l'utilisation dans *GreC* de métadonnées pour lesquelles nous proposons une abstraction au travers d'un métamodèle de GreC.

La suite du papier est organisée de la manière suivante. Nous commencerons par rappeler le cadre général de l'approche *GreC* et monterons comment les concepts classiques de l'OLAP doivent être revisités et étendus (Section 2). Puis nous nous focaliserons sur l'opérationnalisation de l'approche avec l'introduction de métadonnées, de leur modélisation et avec des considérations calculatoires (Section 3). En-

fin nous présenterons l'implémentation d'un prototype pour montrer la faisabilité de l'approche et discuterons des performances (Section 4), pour finalement conclure dans une dernière section (Section 5).

2. Cadre général de l'approche *GreC*

Nous proposons l'approche *GreC* (Graphes enrichis par des Cubes) qui est une nouvelle façon de considérer la combinaison de l'OLAP et des graphes pour l'analyse de réseaux d'information. *GreC* est une approche originale et complémentaire des approches basées sur une construction d'un cube de graphes (Jakawat *et al.*, 2016b). Elle permet de construire un graphe qui réponde aux besoins d'analyse de l'utilisateur et de l'enrichir avec des cubes de données qui vont décrire et valuer les nœuds et/ou les arêtes selon les besoins d'analyse. L'utilisateur peut ainsi avoir une vue globale d'une partie du réseau avec des informations multidimensionnelles et faire des analyses intéressantes en naviguant au sein du graphe enrichi avec des opérateurs dédiés. *GreC* considère la structure du réseau pour permettre des opérations OLAP topologiques, et pas seulement des opérations OLAP classiques et informationnelles. La présence d'une dimension temporelle dans les cubes qui valuent les nœuds et/ou les arêtes permet de rendre compte d'une certaine façon de la dynamique du graphe.

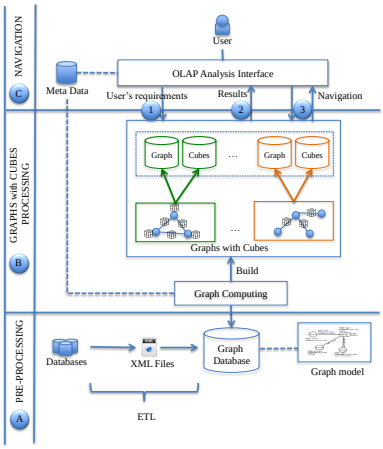


Figure 4. Processus de GreC.

La Figure 4 illustre le fonctionnement global de l'approche *GreC* appliquée, à titre d'exemple, sur les données bibliographiques. Le point de départ est la phase préparatoire (couche A), qui correspond au pré-traitement des données. Diverses bases de données bibliographiques sont fusionnées et intégrées dans des fichiers XML qui alimentent une base de données orientée graphe selon un modèle de données que nous avons défini (Jakawat *et al.*, 2016a). Il s'agit ici de considérer un graphe hétérogène comportant l'ensemble de toutes les données. Ensuite, dans la couche B, à partir du

graphe hétérogène des données de base, les graphes enrichis par des cubes sont construits (notons que l'ensemble des graphes est calculé en amont pour assurer de bonnes performances pour l'utilisateur, nous y reviendrons ultérieurement). Cette construction se décompose en deux étapes : construction du graphe lui-même, puis celle des cubes de données qui valuent les nœuds et/ou les arêtes. Ces deux étapes se répètent pour construire l'ensemble des graphes enrichis par les cubes. Dans la couche C, grâce à une interface de navigation, l'utilisateur exprime ses besoins d'analyse, ce qui permet de sélectionner le graphe adéquat. Une fois le graphe adéquat obtenu, l'utilisateur peut naviguer grâce à des opérateurs adaptés, à la fois par rapport au graphe, mais aussi par rapport aux cubes qui lui sont associés.

Pour permettre l'analyse en ligne de graphes enrichis par des cubes ainsi décrite, nous sommes amenées à redéfinir et à étendre les concepts de l'OLAP manipulés dans le contexte de *GreC*. Tout comme dans l'approche classique d'analyse en ligne, dans *GreC*, il s'agit d'analyser un fait. Par exemple, dans le cadre des données bibliographiques, il peut s'agir d'analyser la production scientifique ou la co-publication. En revanche, le fait n'est pas directement analysé au travers d'une mesure, mais au travers d'un graphe. En fonction du fait, et des besoins d'analyse, des métadonnées permettent de déterminer si des cubes de données valuent des nœuds et/ou des arêtes.

La notion de cube dans *GreC* correspond à un cube classique, qui contient dans chacune de ses cellules la valeur d'une ou plusieurs mesures numériques ; ces mesures peuvent être «simples» (additives) comme le nombre de publications ou elles peuvent être basées sur des graphes comme par exemple une mesure de degré de centralité.

Comme dans l'approche *Graph OLAP* initiale, nous retrouvons deux types de dimension : dimension informationnelle et dimension topologique. Les dimensions informationnelles correspondent aux dimensions définissant les cubes de données attenants aux nœuds ou aux arêtes. Les dimensions topologiques correspondent aux dimensions par rapports aux éléments représentés au niveau du graphe, avec dans les deux cas, la possibilité d'une hiérarchisation. Par exemple, la dimension topologique *auteur* est hiérarchisée avec un niveau *institution*. Ceci permettra de passer du graphe des auteurs au graphe des institutions par exemple. De plus, nous parlons, non pas d'opérateurs OLAP, mais d'opérateurs OLAP informationnels ou topologiques, déterminant ainsi si l'opération (que ce soit un *Roll Up*, *Drill Down*, etc.) est appliquée par rapport au graphe en question selon une dimension topologique, ou aux cubes de ce graphe selon une dimension informationnelle.

Pour rendre opérationnelle l'approche *Grec* sur l'ensemble du processus, nous avons besoin de définir des métadonnées (en plus du modèle de données spécifique à chaque réseau d'information) et de développer de nouveaux algorithmes pour construire les graphes et les cubes, calculer les mesures et adapter les concepts OLAP.

3. Opérationnalisation de GreC

Rendre opérationnelle l'approche GreC pour les données bibliographiques et aussi pour d'autres données nécessite de penser une mise en œuvre assurant la généricité de l'approche. Ceci passe en particulier par la définition de métadonnées.

3.1. Métadonnées pour la généricité

Pour permettre la mise en œuvre de l'approche *GreC*, il est nécessaire que les données de base à considérer soient modélisées dans un graphe hétérogène retraçant l'ensemble des données et leurs liens à prendre en compte. Ce modèle est spécifique à chaque réseau étudié et ne peut faire l'objet d'un modèle générique. Le modèle que nous proposons pour l'exemple de l'analyse des données bibliographiques est introduit dans la section 4.2 lors de la présentation des données pour les expérimentations.

En revanche, d'une façon générale, à partir du graphe hétérogène complet, pour extraire et construire les graphes enrichis par des cubes ainsi que pour assurer la généricité de l'approche *GreC* et notamment celle de l'interface de navigation, nous introduisons des métadonnées avec un modèle dédié (Figure 5), correspondant au métamodèle de *GreC*.

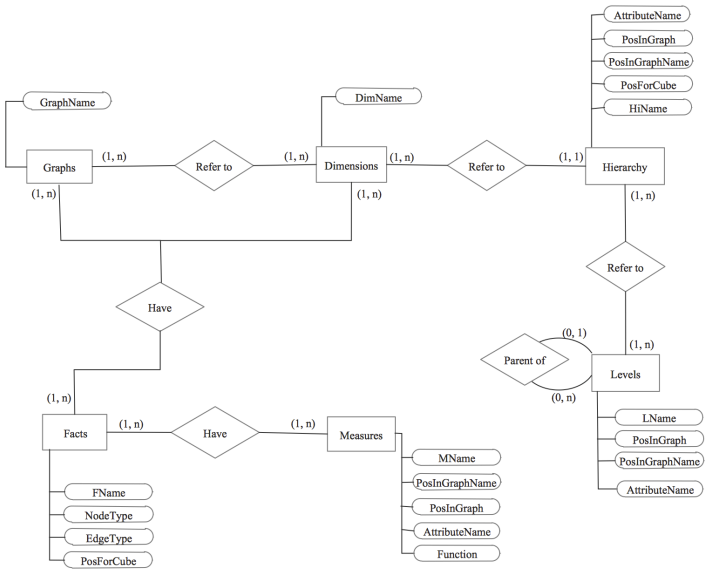


Figure 5. Métamodèle simplifié de GreC.

En adoptant le formalisme du modèle conceptuel entité-association (EA), les principales entités sont les "faits", "mesures", "dimensions", "hiérarchies", "niveaux" et "graphes". Ceci a pour but de représenter les contextes d'analyse possibles pour

l'utilisateur, en partant des faits à analyser et de comment le faire : avec quel graphe, quels cubes, grâce à l'instanciation de ces métadonnées.

Les principales associations permettent alors de déterminer quels graphes sont possibles par rapport à un fait à analyser (association entre *FACTS* et *GRAPHS*) ; quels indicateurs sont possibles (association entre *FACTS* et *MEASURES*). L'entité *DIMENSIONS* est associée à la fois à l'entité *GRAPHS* et *FACTS*, ce qui permet de préciser les dimensions topologiques et informationnelles respectivement, avec ensuite une association avec l'entité *HIERARCHY*, associée elle-même à l'entité *LEVELS*, ce qui permet de déterminer les niveaux de navigation. La définition des mesures et des dimensions informationnelles déterminent la structure du cube.

Chaque entité est bien sûr décrite par différents attributs. Ici, nous mettons en avant les attributs ayant un rôle particulier. Pour l'entité *FACTS*, notons que le fait est précisé au travers du type de nœuds constituant le graphe qui permet d'analyser ce fait (attribut *NodeType*). Chaque fait sera analysé au travers d'un graphe homogène. Nous y trouvons également l'attribut *PosForCube* qui permet de caractériser la position des cubes qui vont enrichir le graphe : au niveau des arêtes, des nœuds, ou à la fois des arêtes et des nœuds. L'attribut *EdgeType* permet de préciser le chemin type dans le graphe initial hétérogène qui permettra la construction du graphe homogène en fonction du fait défini. Au niveau de l'entité *MEASURES*, il est précisé la façon d'agréger les données avec l'attribut *Function*, qui prend par exemple la valeur "numeric" si la mesure peut s'agréger simplement par additivité, ou "degree" s'il s'agit d'une mesure de type calcul de degré basée sur les graphes et nécessitant donc un recalcul de la mesure en fonction des choix d'analyse ou des opérations appliquées, et ce à partir des données sources du graphe initial hétérogène.

Cette modélisation peut s'illustrer sur l'exemple des données bibliographiques en instanciant ces métadonnées :

– Si l'utilisateur veut analyser la co-publication, l'entité *FACTS* permet de tracer le fait que le réseau des co-publications est un graphe où les noeuds sont les auteurs et les arêtes entre deux d'entre eux indiquent qu'ils ont co-écrit ensemble. Cela permet de préciser aussi que ce graphe a des cubes seulement sur les arêtes, puisque l'analyse est centrée ici sur la co-publication et non sur la publication, avec dans ce dernier cas, des cubes qui seraient à la fois sur les arêtes et sur le nœuds, spécifiant que des publications peuvent être écrites par un unique auteur sans collaboration.

– Si le fait est la co-publication, les mesures peuvent être le nombre de papiers, cela peut aussi être une mesure basée sur le graphe comme le degré de centralité. Dans le premier cas, nous avons donc des cubes au niveau des arêtes, mais pour le degré de centralité qui permet d'estimer l'activité d'un nœud, le cube contenant ce type de mesure se trouverait au niveau des nœuds. Ainsi, dans ce dernier cas, il s'agit pour chaque auteur du graphe d'avoir un cube qui caractérise le nombre total de liens (avec des co-auteurs différents) en fonction des dimensions du cube choisies, permettant d'analyser les auteurs actifs dans des collaborations variées.

– Si le fait est la co-publication, plusieurs dimensions comme *time* et *venue* peuvent être utilisées au niveau des cubes qui seront définis. Ainsi, si la mesure est le nombre de papiers pour ce fait, cela induit qu'au niveau des arêtes, il y a des cubes qui déterminent le nombre de papiers co-publiés par année et par conférence.

– Une dimension peut être structurée selon une hiérarchie. Par exemple la dimension institution a une hiérarchie du style : *author name* / *institution name* / *country*, *country* étant un niveau plus élevé de *institution name*. Dans le cadre de l'analyse de co-publication, cette hiérarchie de dimension topologique permettra de faire des opérations pour analyser le phénomène de co-publication à l'échelle des auteurs, mais également des institutions, ou même des pays pour analyser l'internationalisation des collaborations scientifiques.

Le modèle des métadonnées est ainsi conçu pour assurer la généricité de l'approche. Il permet de faire le lien entre les faits à analyser, les graphes à construire et l'emplacement des cubes (au niveau des nœuds et/ou des arêtes). Il permet également de décrire les concepts OLAP (faits, mesures, dimensions, etc.) et de stocker leur instanciation. Les métadonnées sont utilisées par les différents algorithmes qui construisent les graphes et les cubes, qui calculent les mesures et qui réalisent les opérations OLAP redéfinies. Elles conditionnent également l'interfaçage de l'application.

3.2. Considérations calculatoires

Rappelons que le point de départ est un graphe hétérogène avec l'ensemble des données. En fonction des besoins d'analyse exprimés par l'utilisateur, un graphe est proposé à ce dernier, il pourra naviguer dans les données de celui-ci grâce à différents opérateurs OLAP adaptés à l'approche GreC : navigation dans le graphe ou dans les cubes qui enrichissent le graphe.

Différents algorithmes ont été implémentés pour mettre en œuvre *GreC*, en se basant sur l'usage des métadonnées, le graphe initial hétérogène et les besoins d'analyse, cela recouvre notamment les étapes suivantes :

1. la construction du graphe pour l'utilisateur ;
2. la construction des cubes pour valuer les nœuds et/ou les arêtes ;
3. le calcul des mesures numériques (simples ou basées sur les graphes) pour le remplissage des cubes.

En raison des limitations de place, nous ne pouvons donner tous les détails. Pour les mesures numériques classiques qui sont basées sur des comptages, le calcul du cube se fait de façon classique par interrogation des données. Pour le calcul de mesures basées sur le graphe, cela nécessite des parcours de graphes, en fonction du type de la mesure. Par exemple, dans la Figure 6 est représenté le cube contenant la mesure numérique du degré de centralité de l'auteur J. Han en fonction des années et des conférences, et ici en particulier pour EDBT 2009. L'illustration de son calcul pour une cellule est donc basé sur le graphe associé à J. Han, où E1, E4 et E5 correspondent à des arêtes pour lesquelles des papiers à EDBT 2009 sont concernés (d'après les

informations stockées). Il s'agit alors, pour cette mesure de comptabiliser les arcs en question.

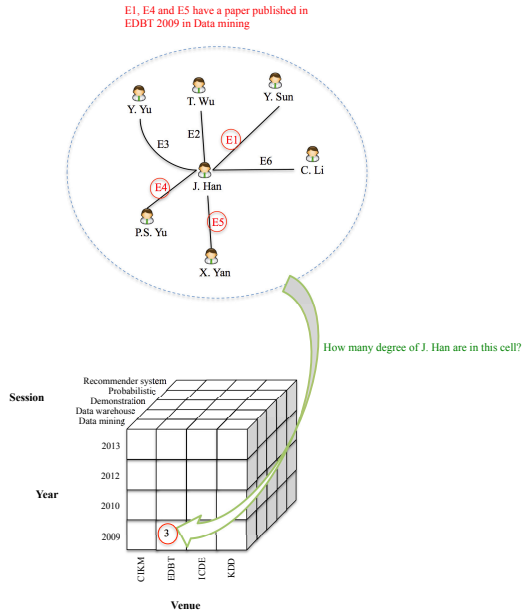


Figure 6. Illustration de cube contenant une mesure numérique calculée à partir du graphe.

Concernant les opérateurs OLAP, ceux-ci sont adaptés aux manipulations du graphe et des cubes. Cela concerne des opérateurs travaillant sur les informations hiérarchiques tels que *Roll Up* et *Drill Down*, mais aussi d'autres opérateurs qui permettent par exemple la sélection de données. Selon les cas, le graphe est amené à changer de structure (type de nœuds) ou les cubes sont recalculés.

Notons que pour éviter certains problèmes d'additivité, le retour aux données sources est souvent nécessaire pour différents calculs lors de la phase de manipulation des cubes ou du graphe, au travers d'une représentation à base de chemins.

Ce retour aux données sources est important d'un point de vue calculatoire pour deux raisons principales. La première est que cela permet de prendre en compte le fait que lorsqu'un *Roll Up* topologique est fait, les résultats demeurent cohérents. Par exemple, prenons le cas de l'analyse des co-publications avec le nombre de papiers comme mesure. Supposons qu'un papier a été écrit par deux auteurs du même établissement, ce papier sera comptabilisé pour les co-publications entre auteurs ; si nous passons au niveau des établissements, ce papier ne sera pas comptabilisé car il ne s'agit pas d'une collaboration inter-établissements. La deuxième raison réside dans le fait que nous prenons en compte l'évolution des données, notamment le fait qu'un auteur peut changer d'établissement dans le temps. Ainsi, nous nous ramenons tou-

jours à la donnée de base qui est la publication. Il s'agit de fait de pouvoir récupérer l'affiliation indiquée pour l'auteur dans le papier en question. Ainsi, l'affiliation d'un auteur qui évolue dans le temps est bien prise en compte dans les différents calculs de graphes et dans les opérations OLAP appliquées.

4. Implémentation et performances

4.1. Caractéristiques du prototype

L'implémentation de GreC a été réalisée en combinant les données bibliographiques de DBLP, ACM et Microsoft Research Area. Ceci est notamment justifié par la complémentarité des données, en termes de récupération des informations sur les affiliations des auteurs des papiers entre autres.

L'architecture de l'implémentation est présentée dans la Figure 7. Les données bibliographiques de base ont été centralisées dans le système NoSQL Neo4j. Les différents graphes correspondant aux différents faits et leurs cubes associés sont générés à partir des données de base du réseau hétérogène et des métadonnées. Les cubes sont ensuite stockés également dans Neo4j.

Les métadonnées sont stockées dans le système relationnel Oracle. Les interfaces pour l'utilisateur ont été développées en Java. Pour assurer la généricité de l'approche, leur contenu est généré en fonction des métadonnées, et des besoins d'analyse exprimés par l'utilisateur.

Concernant la navigation, comme une phase de pré-traitement permet de précalculer les éléments, cette brique consiste à la sélection et la visualisation des données adéquates.

4.2. Présentation des données de base

Pour permettre la mise en œuvre de l'approche *GreC*, les données de base doivent être modélisées. La Figure 8 montre le modèle proposé pour les données bibliographiques. Ce modèle couvre l'ensemble des données considérées et intègrent les liens entre les données à prendre en compte. Ces données de base correspondent à un graphe hétérogène.

4.3. Performances

L'étude de performances développée a permis d'analyser différents points. Elle a été conduite avec Java 1.7.0_75 sur un ordinateur portable avec un processeur Intel core i5 2.4 GHz avec 8 GB de RAM sur Mac OS X version 10.9.2. Quatre jeux de données de tailles différentes ont été constitués avec les caractéristiques décrites dans la Table 1, afin de mesurer l'approche en fonction du volume de données considérées.

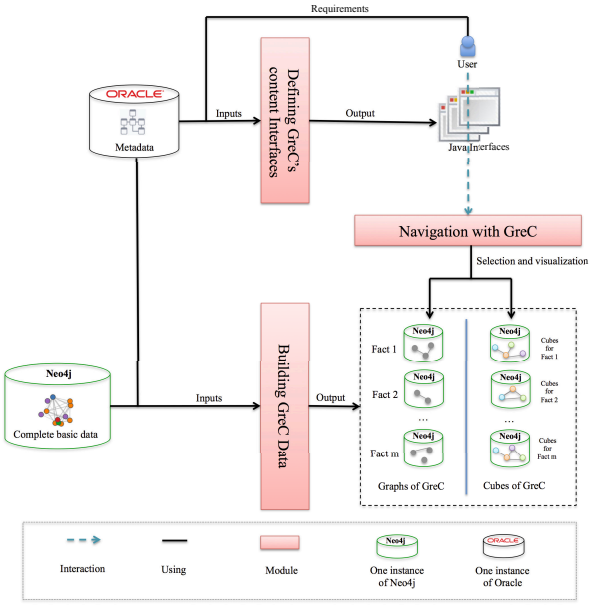


Figure 7. Architecture de l'implémentation de GreC.

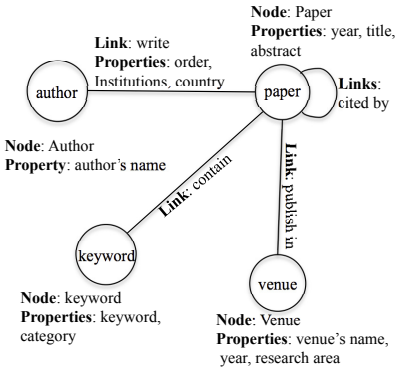


Figure 8. Modèle du graphe des données bibliographiques de base.

Table 1. Jeux de données pour l'expérimentation de GreC.

Jeux de Données	Nb de Publications	Réseau de co-auteurs		Réseau d'institutions	
		Nb de nœuds	Nb d'arêtes	Nb de nœuds	Nb d'arêtes
D1	1000	2216	4322	696	959
D2	2000	3790	8094	1157	1820
D3	3000	5335	12150	1573	2711
D4	4000	7038	16107	2051	3575

Le premier point porte sur la construction du graphe (des différents graphes) pour l'utilisateur. Cet algorithme est une optimisation d'un algorithme existant (Beheshti *et al.*, 2012). La Figure 9 reprend la construction du graphe de co-auteurs (Q1) et celle du graphe représentant les liens de co-publication inter-institutions (Q2). L'étude de performances démontre que sur des jeux de données de plus en plus importants, l'adaptation proposée est plus performante, et cela est possible car le nombre de lectures du graphe initial hétérogène a été optimisé.

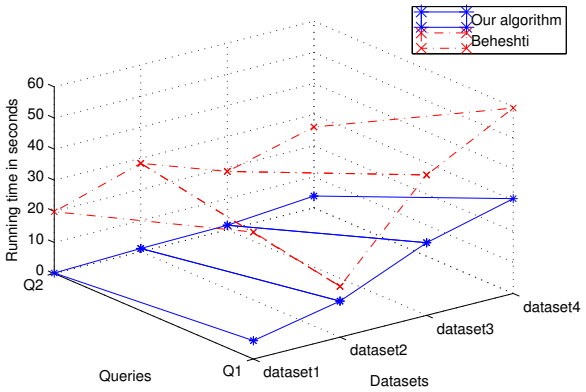


Figure 9. Temps d'exécution de la construction de graphe.

Le deuxième point correspond aux calculs des cubes, et donc a fortiori des mesures. Il apparaît que le temps est raisonnable lorsqu'il s'agit de mesures numériques simples. Le temps augmente fortement lorsqu'il s'agit de mesures basées sur le graphe (qui nécessitent un parcours de graphe pour le calcul), et ce de façon proportionnelle à la taille du graphe. Ceci dépend néanmoins du type de mesure, par exemple, si la mesure se base sur un calcul de voisinage comme le degré de centralité, le temps reste raisonnable comme le montre la Figure 10.

Le troisième point concerne la navigation par l'utilisateur dans les données. Cette étude montre que compte-tenu de l'intérêt d'une mesure à base de graphe en terme analytique et de l'enjeu du temps de traitement, le pré-calcul de ces informations est nécessaire, mais pourrait être partiel selon le contexte . Cette étude montre que compte-tenu des choix de pré-calcul, des temps acceptables de navigation sont obtenus selon les types de requêtes testés et le volume de données. En effet, si les requêtes de type sélective sur des membres de dimension peuvent aller jusqu'à une dizaine de secondes, globalement, le temps d'exécution des requêtes est largement inférieur à 1 seconde pour des requêtes de type *Roll Up* des auteurs aux institutions par exemple. Ainsi, pré-calculer les graphes et les cubes constitue une stratégie intéressante pour optimiser les temps d'exécution pour l'utilisateur.

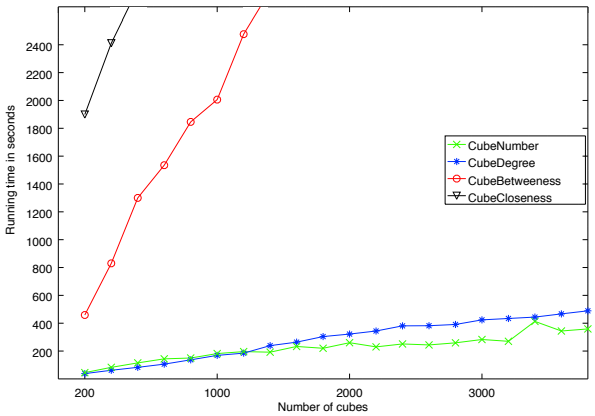


Figure 10. Temps d'exécution de la construction des cubes.

5. Conclusion et perspectives

L'approche GreC proposée constitue une vision innovante et complémentaire dans le domaine du *Graph OLAP*, en permettant à la fois une vue globale du graphe, tout en l'enrichissant d'informations pertinentes et riches au travers de cubes qui valuent les nœuds et/ou les arêtes selon les besoins d'analyse. La présence d'une dimension temporelle dans ces cubes permet d'avoir des éléments sur la dynamique du graphe et de prendre en compte les modifications de données au cours du temps. Nous nous sommes particulièrement focalisés ici sur l'explicitation des éléments qui permettent de donner un caractère générique à cette approche. Nous avons également précisé l'implémentation de *GreC* et donné des éléments de performance qui tendent à montrer la pertinence de l'approche à la fois du point de vue des possibilités d'analyse, mais également de la rapidité de leur obtention selon des choix qui optimisent le temps pour la navigation dans les données.

Ce travail ouvre de nombreuses perspectives. Une première perspective consiste à doublement étendre les possibilités d'analyse offertes par *GreC*. D'une part, il s'agit d'explorer la possibilité d'utiliser des mesures de centralité pour les arêtes. D'autre part, nous voulons introduire dans *GreC* des mesures textuelles. En effet, une grande partie de l'information contenue dans les réseaux d'information est textuelle. L'idée est de combiner les approches *Graph OLAP* aux approches *Text OLAP* afin de proposer une approche complète pour l'analyse des réseaux d'information.

La deuxième perspective concerne l'analyse de l'évolution du graphe. Au-delà de l'aspect temporel des cubes, une piste à explorer serait d'envisager des opérations binaires (différence, intersection, etc.) entre deux graphes issus de *GreC*. Ceci induit de redéfinir ces opérateurs au regard de l'approche *GreC*.

Concernant les données bibliographiques notamment, les publications sont souvent écrites par plus de deux auteurs. Cela pose alors la question de la possibilité

d'avoir recours aux hypergraphes, avec toutes les adaptations qui découleraient de ce choix.

Enfin, il s'agit de se focaliser davantage sur l'utilisateur, avec d'une part développer la possibilité de mieux cerner le graphe ou sous-graphe à analyser (système de recommandation par exemple) et de procéder à une évaluation utilisateur à grande échelle, en terme non seulement d'usage et de performances.

References

Beheshti S.-M.-R., Benatallah B., Motahari-Nezhad H. R., Allahbakhsh M. (2012). A framework and a language for on-line analytical processing on graphs. In *13th International Conference on Web Information Systems Engineering (WISE'12)*, p. 213-227.

Chen C., Yan X., Zhu F., Han J., Yu P. S. (2008). Graph OLAP: Towards online analytical processing on graphs. In *8th IEEE International Conference on Data Mining (ICDM'08)*, p. 103-112.

Jakawat W., Favre C., Loudcher S. (2016a). Graphs enriched by cubes for OLAP on bibliographic networks. *International Journal of Business Intelligence and Data Mining*, Vol. 11, No. 1, pp. 85–107.

Jakawat W., Favre C., Loudcher S. (2016b). OLAP cube-based graph approach for bibliographic data. In *42nd International Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM'16), Student Research Forum.*

Jin X., Han J., Cao L., Luo J., Ding B., Lin C. X. (2010). Visual cube and on-line analytical processing of images. In *19th ACM International Conference on Information and Knowledge Management (CIKM'10).*

Loudcher S., Jakawat W., Morales E. P. S., Favre C. (2015). Combining OLAP and information networks for bibliographic data analysis: a survey. *Scientometrics*, Vol. 103, No. 2, pp. 471–487.

Morfonios K., Koutrika G. (2008). Olap cubes for social searches: Standing on the shoulders of giants? In *International Workshop on the Web and Databases (WebDB).*

Qu Q., Zhu F., Yan X., Han J., Yu P., Li H. (2011). Efficient topological olap on information networks. In *Proceedings of the 16th International Conference on DAtabase Systems For Advanced Applications (DASFAA'11)*, Vol. 1, p. 389-403.

Tian Y., Hankins R., Patel L. (2008). Efficient aggregation for graph summarization. In *ACM SIGMOD International Conference on Management of Data (SIGMOD'08)*, p. 567-580.

Yin M., Wu B., Zeng Z. (2012). Hmgraph olap: a novel framework for multi-dimensional heterogeneous network analysis. In *15th International Workshop on Data warehousing and OLAP (DOLAP'12)*, p. 137-144.

Zhao P., Li X., Xin D., Han J. (2011). Graph cube: On warehousing and olap multidimensional networks. In *ACM SIGMOD International Conference on Management of Data (SIGMOD'11)*, p. 853-864.

Ingénierie des méthodes

Les méthodes d'évolution continue au sein des organisations : le cadre As-Is/As-If

Agnès Front, Dominique Rieu, Ornela Cela, Fatemeh Movahedian

Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, F-38000 Grenoble, France
prenom.nom@univ-grenoble-alpes.fr

RESUME. L'évolution des organisations pour faire face aux nouveaux enjeux socio-économiques nécessite de les doter de méthodes leur permettant de renforcer leurs aptitudes à innover ou co-innover au sein d'écosystèmes complexes, d'améliorer en continu leurs capacités d'absorption de connaissances externes, ou encore d'optimiser leurs processus métier. Pour être efficace, les méthodes proposées doivent être intégrées dans la vie de l'organisation et prises en main par les acteurs, y compris parfois les citoyens eux-mêmes. Le cadre méthodologique As-Is/As-If présenté dans cet article propose un cadre contraint dédié à l'ingénierie des méthodes d'accompagnement des évolutions. Ces méthodes sont caractérisées par des démarches continues intégrant des cycles complets d'évolution et des langages et outils impliquant individuellement et collectivement les acteurs. Le cadre As-Is/As-If est illustré au travers de trois méthodes : ADInnov dédiée à l'innovation dans les écosystèmes, ISEACAP dédiée à améliorer l'absorption des connaissances dans les PME et CEFOP dédiée à l'évolution continue des processus métier des organisations.

ABSTRACT. Evolving organizations in order to face to new socio-economic challenges implies to dote these organizations with news methods allowing them to enrich their capacity to innovate or co-innovate within complex ecosystems, to improve in a continual way their external knowledge absorptive capacity or to optimize their business processes. In order to be efficient, the proposed methods have to be integrated in the organization's life and adopted by the functional actors of the organizations or even the citizens themselves. The methodological framework As-Is/As-If presented in this paper proposes a constrained framework dedicated to the engineering of methods for accompanying evolutions in organizations. These methods are characterized by continuous processes integrating complete evolution cycles, and by languages and tools implying actors in individual and collective activities. This methodological framework is illustrated within three methods: ADInnov dedicated to innovation in ecosystems, ISEACAP dedicated to the improvement of the knowledge absorptive capacity in SME, and CEFOP dedicated to continuous business process evolution in SME.

MOTS-CLES : évolution continue, innovation, écosystème, processus métier, ingénierie des méthodes

KEYWORDS: continuous evolution, innovation, ecosystem, business process, method engineering.

1. Introduction

Pour accompagner l'évolution permanente de notre société et de son économie, de nouveaux enjeux socio-économiques apparaissent obligeant les entreprises à faire face à un marché de plus en plus concurrentiel. Ces nouveaux enjeux engendrent pour les entreprises le besoin de mettre en œuvre des cycles d'amélioration continue de leurs processus métier ou de s'inscrire dans des réseaux collaboratifs afin de développer leurs capacités et compétences en innovation. Parallèlement, l'innovation sociale, organisationnelle et économique se développe au sein d'écosystèmes socio-économiques impliquant la nécessité d'introduire des plateformes d'économie collaborative au sein d'écosystèmes citoyens. Du point de vue de l'ingénierie des méthodes, ces nouveaux enjeux sociétaux et économiques impliquent le besoin de revisiter les méthodes existantes, qu'il s'agisse de méthodes d'amélioration continue, de méthodes de co-conception, ou encore de méthodes de management de l'innovation. En effet, bien que très différents, ces enjeux offrent un cadre contraint commun aux méthodes d'accompagnement qui leur sont dédiés. Répondre à ce cadre constitue aujourd'hui un véritable défi du domaine de l'ingénierie des méthodes. Avant tout, les méthodes proposées doivent être intégrées dans la vie sociale ou professionnelle de l'organisation et prises en main (il ne s'agit pas uniquement de participation) par les acteurs de l'organisation, y compris les citoyens eux-mêmes dans le cas par exemple de plateformes d'économie collaborative. Les méthodes proposées doivent donc :

- reposer sur des cycles d'amélioration **continue**, par opposition aux approches projets ayant une équipe projet, un budget, une date de début et une date de fin, etc. Entrer ou sortir du cycle d'amélioration continue doit être collectivement accepté ;

- permettre aux acteurs de l'organisation d'être le plus **autonomes** possibles et **impliqués collectivement**. En cela les approches participatives bâties sur des langages simples et des outils ludiques doivent être privilégiées. Il s'agit la plupart du temps de remplacer l'expert-méthode par un animateur. Cette propriété est essentielle pour les organisations telles que les PME ou les associations, qui contrairement aux entreprises de grande taille, ne disposent pas d'expert-méthode.

Cet article est destiné à illustrer ces nouveaux enjeux des méthodes au travers d'un nouveau cadre méthodologique appelé Cadre *As-Is/As-If*. La figure 1 présente ce cadre général en utilisant le formalisme MAP[1] (Rolland, 2007). L'approche traditionnelle *As-Is/To-Be* est ici transformée en des cycles itératifs *As-Is/As-If*. L'objectif est d'imaginer des scenarios d'évolution basés sur la question "Et si ?" qui peuvent être déployés à plus ou moins long terme (parfois même très long terme si des évolutions juridiques sont par exemple nécessaires). Les évolutions sont organisées selon des roadmaps spécifiant quand et comment les déployer. La proposition de roadmaps et le déploiement (non détaillés dans cet article) ont pour objectif de prioriser et de dater les évolutions à mettre en place en fonction des

[1] Les modèles MAP sont des graphes dirigés où les nœuds représentent des intentions (buts) et les arcs des stratégies permettant de passer d'un but atteint à un but à atteindre.

contraintes juridiques, économiques, sociales ou techniques impactées par les évolutions proposées. “Sortir” de la méthode doit être un choix collectif des acteurs.

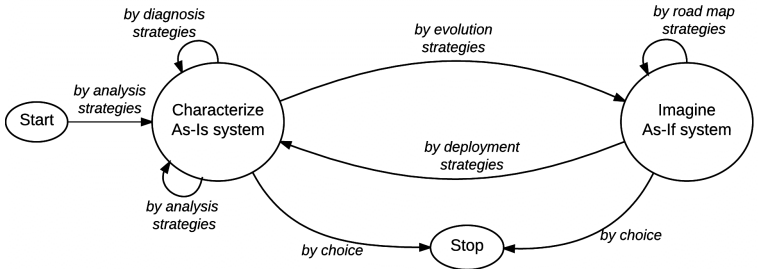


Figure 1 : Cadre méthodologique As-Is/As-If

Ce cadre initialement déployé sous la forme d’une méthode dédiée à l’innovation organisationnelle dans les écosystèmes (Cortes-Cornax et al., 2016) a été repris et déployé dans d’autres contextes : l’amélioration de la capacité d’absorption des connaissances des PME impliquées dans des projets collaboratifs d’innovation (méthode ISEACAP), et l’amélioration des processus métier des PME en prenant en compte l’optimisation interne des processus, mais aussi le déplacement au cours du temps des enjeux du processus (méthode CEFOP). Dans la suite de l’article, la section 2 résume la contribution initiale d’ADInnov (Analysis, Diagnosis, Innovation) pour les innovations humaines et organisationnelles dans les écosystèmes (source du cadre générique). Les sections 3 et 4 introduisent les deux déclinaisons actuelles de ce cadre générique pour la mise en œuvre des méthodes ISEACAP (Identifying, Simulation, Evaluation, Amelioration of Absorptive Capacity) et CEFOP (Continual Evolution of Organisational Process). Enfin, la section 5 présente une conclusion et quelques perspectives.

2. Innovation dans les écosystèmes

Le développement actuel de l’économie collaborative et de l’innovation participative requiert de nouvelles propositions méthodologiques pour aider les acteurs à maîtriser leurs processus d’innovation. Bien que les processus traditionnels d’innovation technologique ou de produits, tels que le « Design Thinking » (Brown, 2009), proposent déjà une approche participative impliquant les utilisateurs de manière précoce dans le processus d’innovation, elles sont aujourd’hui questionnées par ces nouvelles manières d’innover. En effet, l’enjeu n’est plus l’innovation technologique et l’acceptabilité sociale de nouveaux systèmes, produits ou services : les processus d’innovation organisationnelle et sociale tels qu’on les observe aujourd’hui impliquent un écosystème d’acteurs souvent plus complexe au point qu’ils sont parfois pilotés par les citoyens eux-mêmes ou bien par de petites organisations comme des associations, des start-ups, des PME s’associant à des grandes entreprises et / ou des collectivités. A la différence des processus d’innovation technologique, ce nouveau type de processus d’innovation met l’accent

sur des solutions innovantes humaines et organisationnelles où la technologie, et donc parfois l'innovation technologique si elle est utile, vient en support aux écosystèmes collaboratifs (Newmann, 2009) (Debizet, 2016). Les méthodes existantes mobilisées par des modèles de management de l'innovation tels que le Design Thinking ne sont plus adaptées à l'innovation collaborative où des modèles de coordination de services humains sont requis, où la mobilisation de larges communautés est nécessaire et où l'innovation doit être pensée, acceptée et mise en œuvre de manière collective (Bateh et al., 2013). Dans un tel contexte, les écosystèmes entrent dans de nouvelles logiques qui nécessitent l'intégration, par les acteurs, de méthodes permettant de gérer une dynamique d'innovation continue et agile. La méthode ADInnov (Cortes-Cornax et al., 2016a) répond à ce besoin. Elle facilite l'analyse et le diagnostic d'écosystèmes complexes et invite les acteurs de l'écosystème à proposer des innovations sociales, organisationnelles ou techniques consensuelles. ADInnov est issue d'une méthode empirique centrée utilisateur suivie pendant le projet ANR InnoServ[2] dont le but était de trouver des solutions innovantes organisationnelles et techniques pour le maintien à domicile des personnes fragiles.

2.1. Vue générale de la méthode ADInnov

La Figure 2 présente la vue générale de la méthode ADInnov sous la forme d'une MAP avec deux intentions *Characterize As-Is ecosystem* et *Imagine As-If ecosystem* et trois stratégies principales *Analysis*, *Diagnosis* et *Innovation*. La stratégie d'analyse explore le domaine de l'écosystème actuel, identifie ses acteurs et leurs fonctions, et divise l'écosystème actuel en différents réseaux de responsabilité et préoccupations permettant de faciliter la compréhension de la complexité de l'écosystème. La stratégie de diagnostic a pour objectif d'identifier les points de blocage de l'écosystème actuel et d'inférer de façon collective et consensuelle à l'ensemble des acteurs de l'écosystème, des buts permettant de résoudre ces points de blocage. La stratégie d'innovation permet aux acteurs de l'écosystème de proposer des innovations organisationnelles (nouvelles fonctions, nouveaux réseaux de responsabilité, etc.) et des innovations de services ou technologiques permettant de supporter les innovations organisationnelles.

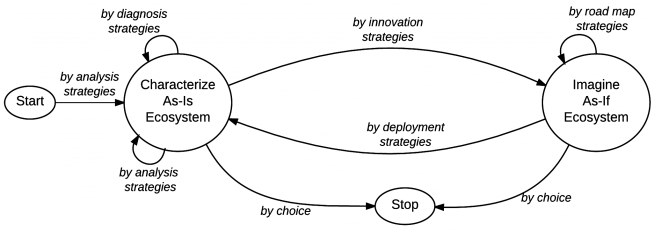


Figure 2 : La méthode ADInnov (Cortes-Cornax et al., 2016a)

[2] https://anrinnoserv.wordpress.com

2.2. Une stratégie particulière

Nous détaillons brièvement ici la stratégie la plus originale de cette méthode, la stratégie d'innovation (les autres stratégies ainsi que le méta-modèle supportant la méthode sont présentés dans (Cortes-Cornax et al, 2016a)). La Figure 3 raffine la section <*Characterize As-Is Ecosystem, Imagine As-If Ecosystem, by innovation strategies*> de la Figure 2. Cette section correspond à la conception des innovations dans l'écosystème *As-Is* de manière à imaginer l'écosystème *As-If*. Les résultats attendus pour cette phase sont un ensemble de services qui aident à atteindre les buts définis dans la phase de diagnostic et un ensemble d'innovations organisationnelles en termes d'altération des acteurs, de leurs fonctions, des réseaux de responsabilité… Dans l'écosystème du projet InnoServ, plusieurs changements organisationnels ont ainsi été proposés, tels que l'introduction du nouveau rôle d'orchestrateur qui réalise la coordination des services autour d'une personne fragile à domicile en utilisant les ressources proches de cette personne. L'introduction de ce nouveau rôle peut potentiellement étendre les prérogatives de certains acteurs, les infirmières pouvant par exemple jouer ce nouveau rôle. Le tableau 1 résume les principales sections mises en évidence en gras dans la Figure 3. Pour chaque section, un exemple issu de l'écosystème du projet InnoServ est donné en italique.

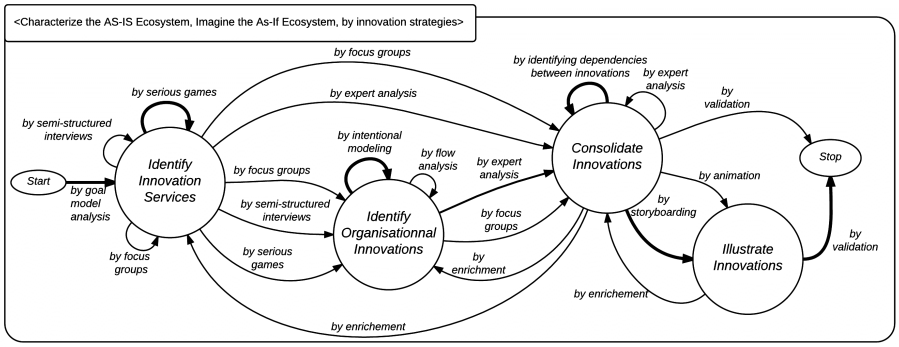


Figure 3 : La stratégie d'innovation de la méthode ADInnov

Tableau 1. Description des sections principales de la stratégie d'innovation

Section	Description
<Start, Identify Innovation Services, **by goal model analysis**>	Identifier les buts permettant d'inférer des solutions concrètes aux points de blocage de l'écosystème. Ex : *Améliorer l'attractivité des métiers du maintien à domicile* permettant de résoudre le point de blocage *Il n'y a pas assez de personnel disponible à domicile.*

<Identify Innovation Services, Identify Innovation Services, **by serious games**>	Utiliser le jeu sérieux Lego Serious Play[3], où les différents participants jouent un rôle (une fonction dans un écosystème) pour proposer des services innovants (voir Figure 4).
<Identify Organizational Innovations, Identify Organizational Innovations, **by intentional modelling**>	Identifier des fonctions qui contribuent à atteindre les buts identifiés. Ex : création du rôle d'*orchestrateur* permettant de coordonner les services autour de la personne fragile à domicile.
<Consolidate Innovations, Illustrate Innovations, **by storyboarding**>	Définir des scénarios d'illustration des services proposés. Ex : scénarios *retour à domicile après hospitalisation* et *gestion de la toilette en environnement rural*.
<Illustrate Innovations, Stop, **by validation**>	Utiliser la méthode CAUTIC (Forest et al. 2013) pour valider les scénarios et les innovations proposées avec les acteurs selon 4 niveaux : Assimilation (relativement au savoir faire technique des acteurs), Intégration (avec les pratiques quotidiennes des acteurs), Appropriation (au regard des rôles et de l'identité des acteurs), Adaptation (à l'environnement des acteurs).

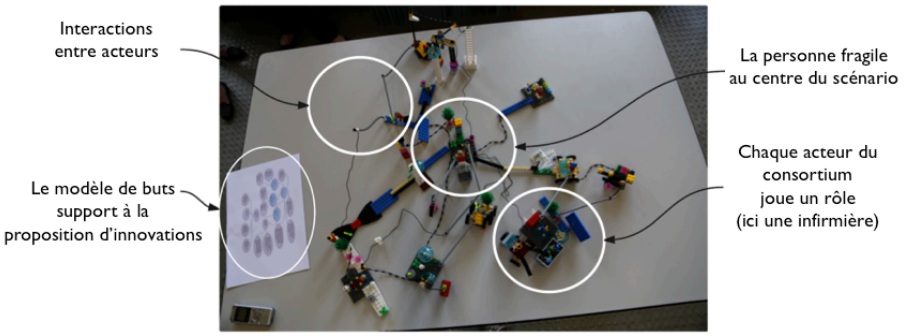


Figure 4 : Exemple de résultat obtenu après la stratégie « by serious game »

2.3. Conclusion

Dans un projet multidisciplinaire comme le projet InnoServ, impliquant des chercheurs de disciplines différentes (juridique, économique, informatique, santé), et où la participation des acteurs même du domaine étudié (maintien à domicile des personnes fragiles) et des associations et acteurs publics et privés liés à ce domaine était indispensable pour que les propositions d'innovations organisationnelles, sociales et techniques aient un sens, il est rapidement devenu évident que les

[3] http://www.lego.com/fr-fr/seriousplay/

méthodes et modes de travail traditionnels n'étaient pas adaptés. De manière totalement empirique, la solution que nous avons alors privilégiée dans le cadre de ce projet a finalement été l'organisation d'ateliers de travail multidisciplinaires de façon à favoriser l'utilisation d'un vocabulaire commun et la richesse d'une confrontation inter-disciplines partagée avec les acteurs du métier. La généralisation de l'approche suivie afin de la rendre applicable à d'autres types d'écosystèmes a donné lieu à la méthode ADInnov. Dans le domaine de l'ingénierie des méthodes, la proposition de cette méthode a engendré de nouveaux enjeux en terme de recherche, puisqu'il convenait en particulier de rendre les acteurs autonomes dans l'utilisation de la méthode, et de combiner au sein de la méthode non plus seulement des fragments de méthodes « classiques », mais également tous types de méthodes issues de domaines multidisciplinaires tels que par exemple les domaines de l'innovation, de la sociologie des usages, ou encore de l'économie expérimentale. Les techniques d'ingénierie de méthodes devront en particulier être à terme revisitées pour prendre en compte cette intégration multidisciplinaire fondamentale dans le champ de l'économie collaborative et participative.

3. Amélioration de la capacité d'absorption des connaissances des PME

Il est aujourd'hui admis que l'aptitude des entreprises à assimiler et utiliser des connaissances externes constitue un facteur clef de réussite en leur permettant d'améliorer leur performance et leur capacité d'innovation (Zollo et al., 2002). Les capacités d'une organisation à acquérir, assimiler, transformer et exploiter les connaissances externes (Nonaka et al., 2000) ont été définies sous le terme de capacité d'absorption (Absorptive CAPacity : ACAP) (Cohen et al., 1990) (Tu et al., 2006). En pratique, elles sont mises en œuvre par des routines organisationnelles (Feldman et al., 2003) qui peuvent être perçues comme des patrons d'activités dont l'exécution répétitive améliore et rationalise l'absorption des connaissances (Zahra et al., 2002) (Becker, 2004).

Dans le cadre du projet ANR ACIC[4], nous sommes partis de l'hypothèse que l'absorption des connaissances externes est particulièrement active dans les projets d'innovation collaborative menés au sein de réseaux d'entreprises, en particulier au sein de réseaux de PME. Renforcer les capacités d'absorption nécessite alors de rationnaliser les processus d'innovation en adoptant un cycle vertueux permettant : 1) d'identifier et tracer les connaissances mobilisées durant un projet d'innovation, 2) en déduire, formaliser et améliorer les routines organisationnelles ayant permis d'identifier, assimiler, transformer et exploiter ces connaissances, 3) intégrer ces routines lors des nouveaux projets d'innovation. Pour supporter ce cycle vertueux, nous avons adopté le cadre méthodologique *As-Is/As-If* dont l'enjeu est ici l'amélioration continue du système (et donc des routines organisationnelles) d'absorption des connaissances. La méthode proposée, appelée ISEACAP, est destinée à aider les organisations à comprendre, améliorer et intégrer de bonnes routines organisationnelles d'absorption des connaissances. Elle permet aux équipes

[4] https://anracic.wordpress.com

projets de jouer ou rejouer leur processus d'innovation en mettant en lumière les pratiques d'ACAP utilisées, pratiques qui peuvent alors être discutées, améliorées et rationalisées au sein de routines organisationnelles à réutiliser lors de nouveaux projets d'innovation. La méthode est basée sur des approches ludiques et participatives qui peuvent être prises en main de manière autonome par les équipes projets. ISEACAP repose sur des séances de modélisation individuelle ou collective mettant en œuvre des techniques d'élicitation des connaissances proposées par (Milton et al. 2007) telles que le « timeline » et le « concept mapping ».

3.1. La méthode ISEACAP

La Figure 5 présente la vue générale de la méthode ISEACAP sous la forme d'une MAP avec deux intentions principales *Characterise As-Is ACAP system* et *Imagine As-If ACAP system*. Les stratégies *by organisation characterisation, by knowledge exploitation, by routine elicitation* constituent des stratégies d'analyse. La stratégie *by evaluating ACAP* est destinée à établir un diagnostic permettant de proposer des stratégies d'amélioration des routines organisationnelles (*by routine amelioration*).

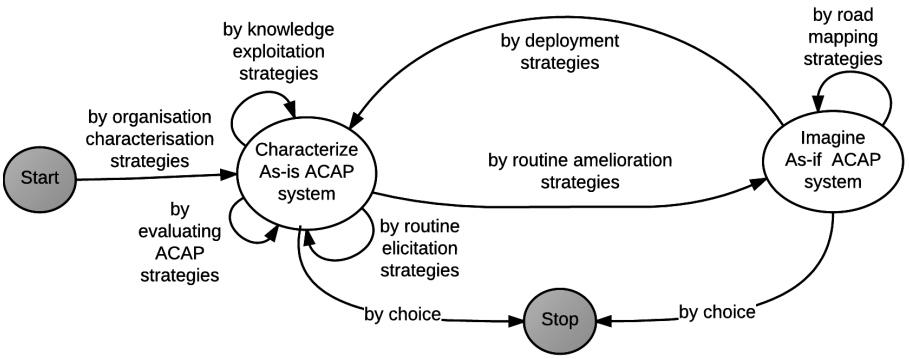


Figure 5 : La méthode ISEACAP

3.2. Une stratégie particulière

La Figure 6 raffine par une MAP la section permettant de caractériser et organiser les connaissances mobilisées durant un projet d'innovation *(<Characterize As-Is ACAP system, Characterize As-Is ACAP system, by knowledge exploitation strategies>*).

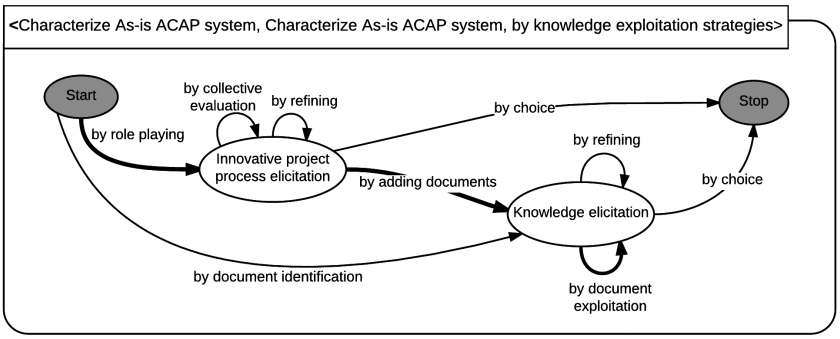


Figure 6 : La stratégie d'exploitation des connaissances de la méthode ISEACAP

La stratégie *by role playing* s'appuie sur la méthode ISEA (Front et al., 2015) et son outil ISEAsy afin de faire rejouer de manière participative un projet d'innovation par ses acteurs afin d'en modéliser les principaux rôles impliqués, étapes menées et documents utilisés pendant ce processus. La stratégie *by adding documents* basée sur un brainstorming, consiste en l'identification des documents les plus intéressants (i.e. porteurs de connaissances externes et riches en connaissances "innovantes"). Certains documents identifiés lors de la modélisation du processus sont retenus. La Figure 7 raffine la stratégie *by document exploitation* et le tableau 2 décrit les sections usuellement suivies dans cette stratégie.

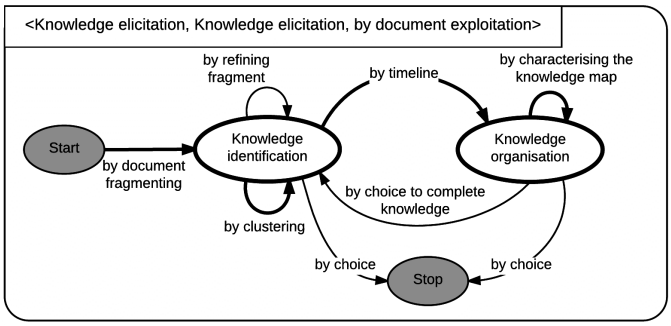


Figure 7 : La stratégie d'exploitation des documents

Tableau 2. Sections principales de la stratégie d'exploitation des documents

Section	Description
<Start, Knowledge identification, **by document fragmenting**>	Chaque participant choisit un document et dispose d'une dizaine de minutes pour en extraire 5 fragments. La question posée est "quelles sont les parties du document qui vous semblent particulièrement importantes pour l'innovation et la bonne conduite du projet ? ». Pour chaque fragment les acteurs décrivent l'information contenue dans

	le fragment.
<Knowledge identification, Knowledge identification, **by clustering**>	Collectivement les acteurs groupent les fragments en se basant sur la proximité des informations contenues et nomment la connaissance obtenue.
<Knowledge identification, Knowledge organisation, **by timeline**>	Les participants organisent les connaissances obtenues sur un axe temporel.
<Knowledge organisation, Knowledge organisation, **by characterising the knowledge map**>	Les participants caractérisent les connaissances (interne/externe, spécifique/générique) et organisent les connaissances : lien de transformation entre connaissances spécifiques, lien de mobilisation des connaissances génériques lors des transformations (exemple figure 8).

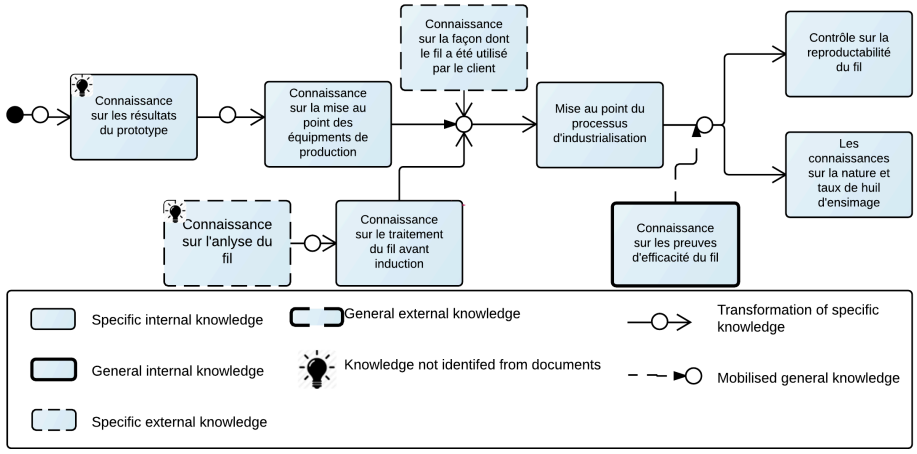


Figure 8 : Cartographie des connaissances

La cartographie des connaissances est ensuite exploitée pour identifier les routines organisationnelles d'ACAP. Elles sont obtenues en « ouvrant » les nœuds de transformation (ronds des liens de transformation, stratégie non présentée ici).

3.3. Conclusion

Le projet ACIC est un projet interdisciplinaire (informatique, gestion, génie industriel). Outre les problèmes de terminologie, c'est surtout l'absence de référentiel méthodologique qui était un frein à l'avancement du projet. L'adoption du cadre méthodologique *As-Is/As-If* a permis d'avancer et d'intégrer des fragments méthodologiques issus de disciplines différentes. Les stratégies décrites ci-dessus ont été pilotées par les informaticiens, alors que la section de diagnostic (non décrite) a été guidée par les gestionnaires. Les approches ludiques et participatives constituent également un moyen de faire fonctionner l'équipe : nous jouons ensemble avant de faire jouer des entreprises. C'est à partir de ces jeux collectifs que

nous avons pu formaliser la démarche mais également le langage (méta-modèle et notation). Il est également important de ne pas partir d'une page blanche : adopter ISEA et son outil ISEAsy nous a permis d'avoir très rapidement des études de cas communes issues de projets réels et de fidéliser pour la suite des entreprises beta-testeurs.

4. L'amélioration continue des processus

Les organisations vivent aujourd'hui dans des environnements concurrentiels qui nécessitent des évolutions stratégiques, organisationnelles et opérationnelles constantes. Ces évolutions doivent être prises en compte dans les processus métier des organisations. Or l'adaptation des processus métier reste un problème constant en terme de temps, de coût mais également de freins internes. Ici aussi l'approche par projet, même itérative et incrémentale, a ses limites : l'esprit d'évolution doit être permanent, vivre au sein de l'organisation et être à la main des acteurs des organisations. Ces hypothèses ont un certain nombre de conséquences méthodologiques : 1) la méthode doit assurer une bonne couverture fonctionnelle permettant l'analyse et le diagnostic du processus, l'identification des évolutions pour répondre à des points de blocage, l'établissement de roadmaps de déploiement des évolutions… 2) l'autonomie des acteurs de l'organisation vis à vis d'une expertise externe est importante pour les organisations n'ayant pas en interne de compétences méthodologiques 3) la motivation et l'implication des acteurs doit joindre l'utile à l'agréable : une réflexion sur les solutions possibles d'évolution des processus doit également favoriser la communication et les relations interprofessionnelles 4) la méthode doit faire face à tout besoin d'évolution et donc aux deux grandes natures d'amélioration des processus : l'amélioration des résultats de l'organisation et celle de ses processus internes. L'ensemble de ces besoins constitue un nouvel enjeu dans les méthodes de gestion des processus métier qu'actuellement aucune méthode n'est en mesure de couvrir. Cette affirmation est issue de notre analyse des méthodes issues de deux domaines principaux : Business Process Improvement (BPI) (Roseman, 2010) (Adnan et al., 2014) et Process Mining (PM) (Van der Aalst 2011). En effet, si la plupart des méthodes issues du BPI (Gershon, 2010) (Pyzdek, 2003) (Zagloe et al, 2011) assurent généralement une bonne couverture fonctionnelle allant de l'analyse et du diagnostic du processus actuel jusqu'à la proposition d'évolutions organisationnelles ou techniques, la conduite de la méthode est effectuée par un expert, dont le rôle est souvent crucial même si la plupart du temps en interaction avec les managers de l'organisation. De plus, ces méthodes sont essentiellement axées sur l'amélioration des résultats du processus et non sur son amélioration interne. Les méthodes basées sur l'approche *Lean* (par exemple Kaizen (Titu et al, 2010) changent au contraire le focus de l'objectif de l'amélioration en s'intéressant à l'amélioration du processus lui-même, en particulier de son délai d'exécution et de son coût. Enfin, la méthode PM2 (van Eck et al., 2015), méthode à notre connaissance la plus aboutie du domaine du PM, a pour objectif l'amélioration du processus interne en diagnostiquant ses points de blocage par l'analyse des traces informatiques liées à l'utilisation des logiciels métier. L'expert intervient en guidant les utilisateurs jusqu'à l'identification des

points de blocage. Leur résolution et les propositions d'améliorations ne font cependant pas partie du périmètre de la méthode. La méthode CEFOP présentée dans la section suivante répond aux quatre besoins évoqués plus haut afin de promouvoir la prise en main par les acteurs des organisations, en particulier des PME, d'une méthode permettant l'amélioration continue des processus métier.

4.1. CEFOP pour l'amélioration continue et autonome des processus métier

La Figure 9 présente la vue générale de la méthode CEFOP sous la forme d'une MAP avec deux intentions *Characterise As-Is Process* et *Imagine As-If Process*. La stratégie *by analysis strategy* permet d'identifier le modèle du processus actuel ainsi que l'objectif principal lié à son évolution. La stratégie *by diagnostic strategy* permet d'identifier ou de prédire des points de blocage dans le processus actuel. La stratégie *by Change Strategy* est destinée à établir des propositions de changements organisationnels ou technologiques à déployer dans le processus *As-If*, qui pourra lui-même être analysé via la stratégie *by Analysis Strategy*.

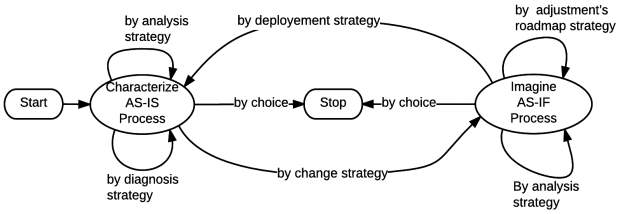


Figure 9 : La méthode CEFOP

4.2. Une stratégie particulière

La Figure 10 raffine la section <*Characterize As-Is Process, Characterize As-Is Process, by analysis strategy* > permettant d'analyser le processus et les objectifs principaux de son amélioration.

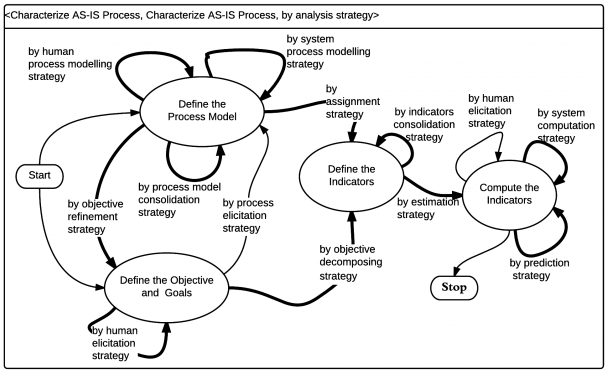
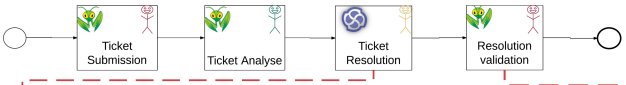


Figure 10 : La stratégie d'analyse de la méthode CEFOP

Tableau 3. Sections principales de la stratégie d'analyse de la méthode CEFOP

Section	Description
<Define the Process Model, Define the Process Model, **by human process modeling strategy**>	S'appuie sur la méthode ISEA (Front et al., 2015) et son outil ISEAsy afin de faire expliciter de manière participative par les acteurs un modèle représentant les rôles et activités du processus organisationnel actuel.
<Define the Process Model, Define the Process Model, **by system process modeling strategy**>	S'appuie sur l'outil de Process Mining Pro-M pour construire, de manière automatisée, un modèle BPMN du processus actuel en fonction des traces laissées par les acteurs du processus actuel dans les outils informatiques.
<Define the Process Model, Define the Process Model, **by process model consolidation strategy**>	Permet la confrontation des deux modèles ISEA et BPMN afin que les acteurs se mettent d'accord sur les principales activités concernées par l'amélioration (voir Figure 11, activités *Ticket submission, Ticket analysis*, etc).
<Define the Process Model, Define the Objective and Goals, **by objective refinement strategy**>	Permet d'identifier l'objectif principal de l'amélioration du processus et sa décomposition en buts. Ex : « *Accélérer la résolution des tickets*» *décomposé en : "Diminuer le nombre de tickets soumis", et "Réduire le temps de résolution d'un ticket"*
<Define the Process Model, Define the Indicators, **by assignment strategy**>	Fait éliciter par les acteurs des indicateurs assignés aux principales activités du processus afin d'évaluer ou de monitorer leur propre performance.
<Define the Objective and Goals, Define the Indicators, **by objective decomposing strategy**>	Fait éliciter par les acteurs des indicateurs contribuant à l'évaluation des buts de l'amélioration.
<Define the Indicators, Define the Indicators, **by indicators consolidate strategy**>	Met en relation les indicateurs avec les activités à évaluer et les buts à atteindre.
<Define the indicators, Compute the indicators, **by estimation strategy**>	Une première estimation des valeurs des indicateurs est réalisée par les acteurs des processus en fonction de leur propre expérience (voir Figure 11).
<Compute the Indicators, Compute the Indicators, **by system computation strategy**>	Les valeurs des indicateurs sont calculées à partir des traces générées par les outils informatiques. Les dates de valeurs de ces indicateurs sont estimées en fonction du temps présent *T0* et passé, par exemple *T0-6 mois, T0-1an*, etc. (voir Figure 11).
<Compute the Indicators, Compute the Indicators, **by prediction strategy**>	En utilisant des outils de business intelligence, calcule une prédiction de la progression des indicateurs dans le futur, par exemple à *T0+6 mois*, *T0+1an*, etc. (voir Figure 11).

La Figure 11 montre un exemple de tableau d'indicateurs issu de la stratégie d'analyse de la méthode CEFOP pour le processus « Gestion des tickets MANTIS ».



Indicator	Values of Indicators			
	Human Values	Now	+ 6 Months	+ 1 Year
Average Number of Days/Ticket Resolution	5	**9**	10,5	12
Maximal Number of Days/Ticket Process Resolution	1 Year	**714**	826	936
Average Number of Days/Ticket Analysis Activity	4	**4**	4.5	5
Maximal Number of Days/Ticket Resolution Activity	3 Month	**642**	758	874
Maximal Number of Days/Ticket Resolution Validation Activity	1 Month	**570**	690	810

Figure 11 : Un tableau d'indicateurs résultat de la stratégie d'analyse

4.3 Conclusion

La méthode CEFOP est issue d'un partenariat avec une start-up soucieuse de l'amélioration de ses processus internes, mais qui ne dispose pas des moyens humains et financiers nécessaires à cette préoccupation. L'utilisation du cadre *As-Is/As-If* combinée à l'exploitation maximisée des traces laissées par les acteurs de l'organisation permet à l'entreprise de prendre conscience des problèmes liés à ses processus et de prédire par le biais d'indicateurs et à moindre coût, l'efficacité de solutions organisationnelles à plus ou moins long terme. En effet, le principal intérêt de la méthode CEFOP n'est pas tant, comme dans les méthodes ADInnov ou ISEACAP, dans le fait qu'elle est destinée à un consortium d'acteurs fonctionnels potentiellement issus de disciplines différents ou non spécialistes en informatique, mais plus dans la confrontation permanente entre des préoccupations et des ressentis humains, et des « calculs » automatisés rendus possible grâce à des outils de process mining ou de business intelligence basés sur les traces réelles laissées par les acteurs. De plus, le calcul de prédictions d'indicateurs est constamment mis en exergue avec les prédictions humaines. La méthode CEFOP est en cours de consolidation, elle est basée sur un méta-modèle et un langage qui seront présentés dans un article ultérieur.

5. Conclusion et perspectives

As-Is/As-If offre un cadre méthodologique facilitant le développement de méthodes d'accompagnement continu au sein des organisations. Il s'applique lorsque le concept traditionnel de projet (équipe projet, délai, coût) n'est pas adapté et lorsque la méthode doit s'intégrer dans la vie de l'organisation (PME, écosystème, etc.). Les méthodes issues de ce cadre sont caractérisées par 1) un cycle d'amélioration continue proposant des stratégies d'analyse, de diagnostic, d'évolution, de planification des évolutions et de déploiement du système étudié 2)

des langages et outils simples et ludiques facilitant une implication collective et autonome des acteurs. Dans cet article nous avons surtout présenté et instancié le cycle *As-Is/As-If*, les aspects langages n'ont été que partiellement illustrés. Mais de même que pour les démarches, le cadre fournit également un cadre de construction du méta-modèle représentant le système étudié. Tous les méta-modèles incluent des concepts décrivant les enjeux de l'évolution du système, ce qu'il est, ses points de blocage, les évolutions envisagées, ainsi qu'une roadmap de ces évolutions. Le méta-modèle n'est complet que si tous ces éléments sont présents (ou absents à dessein).

Une méthode dédiée aux acteurs des organisations ne peut qu'être construite avec eux. Nous avons, dans les trois cas présentés dans cet article, adopté la démarche centrée utilisateur proposée dans (Mandran et al., 2013) (Cortes-Cornax et al., 2016b), démarche incluant tout au long du développement de la méthode des cycles d'évaluation en trois phases : exploration, co-construction et validation. Les langages, démarches et outils sont donc tout au long de leur développement co-conçus et validés avec leurs futurs utilisateurs. Enfin, l'autonomie des organisations passe aussi par l'accès et la formation aux outils support. Pour cela, les outils développés dans le contexte du cadre *As-Is/As-If* seront progressivement intégrés dans la plate-forme **MethodForChange** (https://methodforchange.com/) destinée à héberger une suite de méthodes dédiées à faciliter les projets d'innovation. Ces méthodes sont issues de recherches interdisciplinaires en innovation conduites à l'Université Grenoble Alpes. La plateforme est à l'initiative et est gérée par la fédération de recherche INNOVACS (http://innovacs.univ-grenoble-alpes.fr/).

Bibliographie

Adnan O., Nazir Ahmad M. (2014) *Business Process Improvement Methodologies: An Overview*, Journal of Information systems research and Innovation.

Bateh J., Castaneda M. E., Farah J. E. (2013) Employee Resistance To Organizational Change, Int. J. Manag. Inf. Syst., vol. 17, no. 2, pp. 113–116.

Becker M-C. (2004). Organizational routines: a review of the literature, Industrial and Corporate Change, vol. 13, no. 4, pp. 643-678.

Brown T. (2009), Change by Design: How Design Thinking Transforms Organizations and Inspires Innovation, Broché.

Cohen & Levinthal (1990). Absorptive Capacity: a New Perspective on Learning and Innovation," Administrative Science Quarterly, vol. 35, no. 1, pp. 128-152.

Cortes-Cornax M., Front A., Rieu D., Verdier C., Forest F. (2016a). ADInnov : an Intentional Method to Instill Innovation in Socio-Technical Ecosystems, CAISE 2016, International Conference on Advanced Information Systems Engineering, 15-17 Juin 2016, Slovénie.

Cortes-Cornax M., Dupuy-Chessa S., Mandran M. , Rieu D. et Front A. (2016b) Illustration d'une démarche centrée utilisateur pour l'ingénierie informatique : conception de langage de modélisation, Innovatio (Revue pluridisciplinaire en innovation), no 4. http://innovacs-innovatio.upmf-grenoble.fr/index.php?id=332

Debizet G. (2016). Scénarios de transition énergétique en ville : Acteurs Régulations Technologies. La Documentation française, 978-2-11-010025-2.

Feldman M-S., Pentland, B-T. (2003) Reconceptualizing Organizational Routines as a Source of Flexibility and Change, Administrative Science Quarterly, vol. 48, pp. 94-118.

Forest F., Mallein P., Arhippainen L. (2013) Paradoxical User Acceptance of Ambient Intelligent Systems: : Sociology of User Experience Approach," in Proceedings of International Conference on Making Sense of Converging Media, pp. 211–218.

Front A., Rieu R., Santorum M., Movahedian F. (2015). A participative end-user method for multi-perspective business process elicitation and improvement, Software and Systems modeling, vol. Theme Section Paper, pp. 1-24.

Gershon M. (2010) Choosing which process improvement methodology to implement, Journal of Applied Business and Economics, Vol. 10, No. 5.

Mandran N., Dupuy-Chessa S., Front A. and Rieu D. (2013) "Démarche centrée utilisateur pour une ingénierie de langages de modélisation de qualité," Rev. RTIS-ISI, vol. 18, no. 3

Milton N-R. (2007) Knowledge Acquisition in Practice. A step-by-step Guide, London: Springer-Verlag London Limited.

Newman M. E. J. (2009) Complex Systems: A Survey, Phys. Rep., vol. 79, no. I.

Nonaka I., Toyama R., Konno N. (2000). SECI, Ba and Leadership: A Unified Model of Dynamic Knowledge Creation, Elsevier, pp. 5-34.

Pyzdek T. (2003) The Six Sigma Handbook, Revised and Expanded", ISBN 0-07-141015-5.

Rolland C. (2007), Capturing System Intentionality with Maps, Conceptual Modelling in Information Systems Engineering, pp. 141-158.

Rosemann M. (2010) Handbook on Business Process Management, International Handbooks on Information Systems, Springer, 2010

Titu M., Oprean C., Grecu D. (2010) Applying the Kaizen Method and the 5S Technique in the Activity of Post-Sale Services in the Knowledge-Based Organization, Lecture Notes in Engineering and Computer Science.

Tu Q., Vonderembse M-A., Ragu-Nathan T-S., Sharkey T-W . (2006). Absorptive capacity: Enhancing the assimilation of time-based manufacturing practices, Journal of operations management, vol. 24, no. 5, pp. 692-710.

van Eck M. L., Lu X., Leemans S. J.J., Van der Aalst W. M. P (2015) PM2: a Process Mining Project Methodology, RCIS.

Van der Aalst W. M. P. (2011) Process Mining Discovery, Conformance and Enhancement of Business Processes ", ISBN 978-3-642-19344-6.

Zagloe Z.T., Dachyar M., Nur Arfiyanto F. (2011) Quality Improvement Using Model-Based and Integrated Process Improvement (MIPI) Methodology.

Zahra S-A., George G. (2002). Absorptive Capacity: A review, Reconceptualization and Extension, Academy of Managment Review, pp. 185-203.

Zollo M., Reuer J-J., Singh H. (2002). Interorganizational routines and performance in strategic alliances, Organization Science, vol. 13, no. 6, pp. 701-713.

DMN (Decision Model and Notation) :

De la Modélisation à l'Automatisation des Décisions

Thierry BIARD[1], Jean-Pierre BOUREY[2], Michel BIGAND[2]

1. Laboratoire Génie Industriel - CentraleSupélec – Université Paris-Saclay
Grande Voie des Vignes, 92290 Châtenay-Malabry, France
thierry.biard@centralesupelec.fr

2. École Centrale de Lille - Université Lille Nord de France
Cité Scientifique - CS 20048, 59651 Villeneuve d'Ascq Cedex, France
jean-pierre.bourey@centralelille.fr, michel.bigand@centralelille.fr

RESUME. Cet article s'intéresse à la nouvelle notation DMN (Decision Model and Notation) qui est utilisée pour modéliser de façon standard les prises de décisions. Après une présentation du contexte, les principaux éléments de DMN sont montrés, puis utilisés dans une étude de cas. Les trois modèles CIM, PIM et PSM de la MDA (Model Driven Architecture) sont rappelés, avant d'être appliqués à DMN. Cette approche permet de comprendre, puis de mettre en œuvre deux solutions techniques différentes pour automatiser les prises de décision, but de cet article. Un chapitre est consacré au changement de paradigme : la programmation déclarative (versus la programmation procédurale). Les propos sont illustrés par du code informatique extrait de notre démonstrateur.

ABSTRACT. This article focuses on the new DMN notation (Decision Model and Notation) which is used to model with a standard way the decision-making. After a context presentation, the main DMN elements are shown, then used into a case study. The three models CIM, PIM and PSM of the MDA (Model Driven Architecture) are reminded, before being applied to DMN. This approach allows to understand then to enforce a couple of different technical solutions to automate decision-making, goal of this article. A chapter is dedicated to the paradigm change: the declarative programming (versus the procedural programming). The talking is illustrated by some computer code extracted from our demonstrator.

MOTS-CLES : DMN, décision, modélisation, notation, règle métier, table de décision, automatisation, processus, BRMS, Drools.

KEYWORDS: DMN, decision, model, notation, business rules, decision table, automation, process, BRMS, Drools.

1. Introduction

Dans un contexte général de modélisation d'entreprise, il apparait que les notations standards utilisées pour la représentation des processus métier sont peu adaptées pour la représentation des prises de décisions opérationnelles. Quelques notations intéressantes, à l'initiative d'éditeurs spécialisés, sont apparues ces dernières années, mais elles étaient propriétaires.

Depuis trois ans, une nouvelle notation nommée DMN (Decision Model and Notation) est venue compléter le bouquet de standards proposés par l'OMG (Object Management Group) dans le cadre de la modélisation métier.

Cette notation DMN permet de modéliser les prises de décision et les règles métier. Son objectif principal est de fournir une représentation de la prise de décision compréhensible par toutes les parties prenantes (acteurs métier qui gèrent et pilotent les décisions, analystes métiers qui spécifient les exigences relatives aux prises de décisions, développeurs responsables de l'automatisation de tout ou partie de la prise de décision).

L'aspect visuel de cette notation DMN, qui ne comporte que quatre éléments graphiques principaux, semble trop simple de prime abord pour permettre l'automatisation des prises de décisions qui ont été modélisées. Le but de cet article est de démontrer qu'au contraire, il est possible d'automatiser les prises de décisions qui ont été modélisées avec la notation DMN. C'est sa contribution principale.

Après avoir positionné la notation DMN par rapport aux autres spécifications standards publiées par l'OMG, les principaux éléments de DMN (diagramme et table de décision) seront présentés. La complémentarité entre la notation DMN et la notation BPMN sera évoquée.

La MDA (Model Driven Architecture), qui applique le principe de Séparation des Préoccupations grâce à ses trois modèles CIM, PIM et PSM, va nous fournir la grille de lecture adéquate pour comprendre, puis mettre en œuvre deux solutions techniques différentes, dites spécifique et générique, pour automatiser les prises de décisions modélisées selon la notation DMN.

Ces deux solutions, réalisées à l'aide des mêmes outils gratuits (pour le monde académique au moins) seront succinctement comparées à partir de la même étude de cas, qui consiste à décider (ou non) de la recevabilité d'un vacataire lors d'un processus de recrutement dans un établissement public d'enseignement supérieur.

Sans surprise, ces deux solutions donneront, à partir du même modèle, le même résultat - une prise de décision identique, mais avec deux approches assez différentes. Le critère principal de comparaison sera celui de l'indépendance aux outils utilisés.

Enfin, la conclusion indiquera si le but de cet article est atteint, puis recommandera une solution technique, avant de présager quelques perspectives.

2. DMN (Decision Model and Notation)

2.1. Présentation du contexte

La notation DMN (Decision Model and Notation) [OMG, 2016] est un nouveau standard de l'OMG (Object Management Group). L'OMG publie des dizaines de spécifications standards, dont la plus connue est UML (Unified Modeling Language) [OMG, 2015c]. La Figure 1 positionne les standards qui nous intéressent ici.

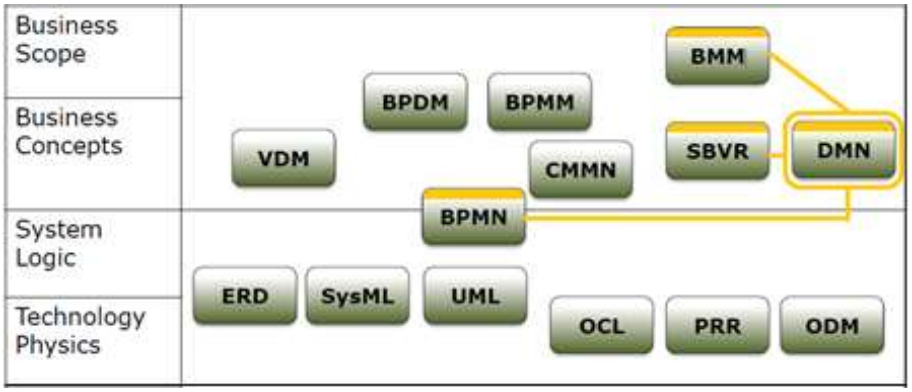


Figure 1: Bouquet de standards de l'OMG, classés selon [Pitschke, 2014]

La notation DMN fait partie des spécifications Business Modeling, avec :

– BMM (Business Motivation Model) [OMG, 2015a], peu connue, car peu utilisée directement, mais qui sert de référence aux autres standards ; une décision représentée en DMN peut supporter un objectif représenté en BMM ;

– SBVR (Semantics of Business Vocabulary and Business Rules) [OMG, 2015b], dont DMN peut utiliser les termes métier qui auront été précédemment définis selon son formalisme [Linehan et de Sainte Marie, 2011] ;

– BPMN (Business Process Model and Notation) [OMG, 2013] ; la notation DMN peut s'utiliser seule ou en complément d'un processus métier représenté en BPMN ; leur complémentarité sera démontrée plus loin.

La première version 1.0 bêta de la spécification DMN a été publiée en février 2014, mais présentait plusieurs lacunes dans son métamodèle (des espaces incongrus dans les noms de classes, voire pas de nom du tout, par exemple), rendant son utilisation difficile. Nous avons néanmoins réussi à développer sur cette base notre propre outil de modélisation DMN [Biard et al., 2015].

C'est la version 1.1, publiée en juin 2016, qui a permis enfin son utilisation réelle et qui sert de base à la plupart des outils disponibles sur le marché aujourd'hui. DMN est donc une notation standard assez récente et encore méconnue.

2.2 Principaux éléments de DMN

2.2.1. DRD (Decision Requirements Diagram)

La représentation emblématique de la notation DMN est le *Decision Requirements Diagram* (DRD). Ce diagramme contient quatre éléments graphiques principaux. Les deux premiers éléments sont obligatoires ; les deux derniers facultatifs :

– *Input Data* (Donnée d'entrée ; au moins une, généralement plusieurs) ;

– *Decision* (Décision ; au moins une ; la Décision est en fait la donnée de sortie).

– *Knowledge Source* (Source de connaissance ; un expert ou un document de référence qui fait autorité dans le métier concerné) ;

– *Business Knowledge Model* (Modèle de connaissance métier, contenant une logique de décision, qui peut être réutilisé dans plusieurs diagrammes DRD).

Trois types de liens différents, les *Requirements*, permettent de représenter les relations entre ces quatre éléments graphiques principaux :

– *Information Requirement*, trait plein terminé par une pointe, qui relie :

- soit une Donnée d'entrée vers une Décision ;

- soit une Décision secondaire vers une Décision principale ;

– *Authority Requirement*, trait pointillé terminé par un point, qui symbolise l'invocation d'une Source de connaissance ;

– *Knowledge Requirement*, trait pointillé terminé par une pointe, qui symbolise l'invocation d'un Modèle de connaissance métier.

Tous ces composants graphiques de DMN (les quatre éléments principaux et les trois liens différents) sont utilisés dans le diagramme DRD de la Figure 2 :

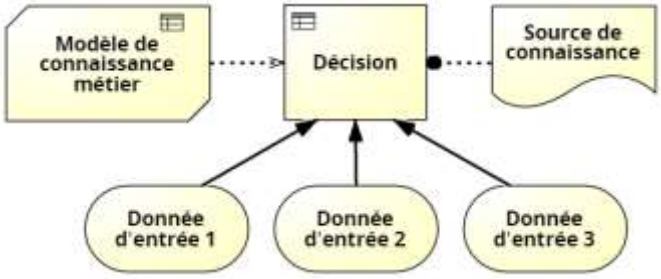


Figure 2: Tous les composants graphiques de DMN dans un diagramme DRD

Les types de liens adéquats pour relier les quatre éléments principaux sont automatiquement sélectionnés par les outils de modélisation, conformément au métamodèle de DMN, fourni par l'OMG. Ces liens sont explicites dans ce métamodèle (extrait en Figure 3), c'est-à-dire représentés par des classes, ce qui n'est pas très courant (les liens-relations sont généralement implicites).

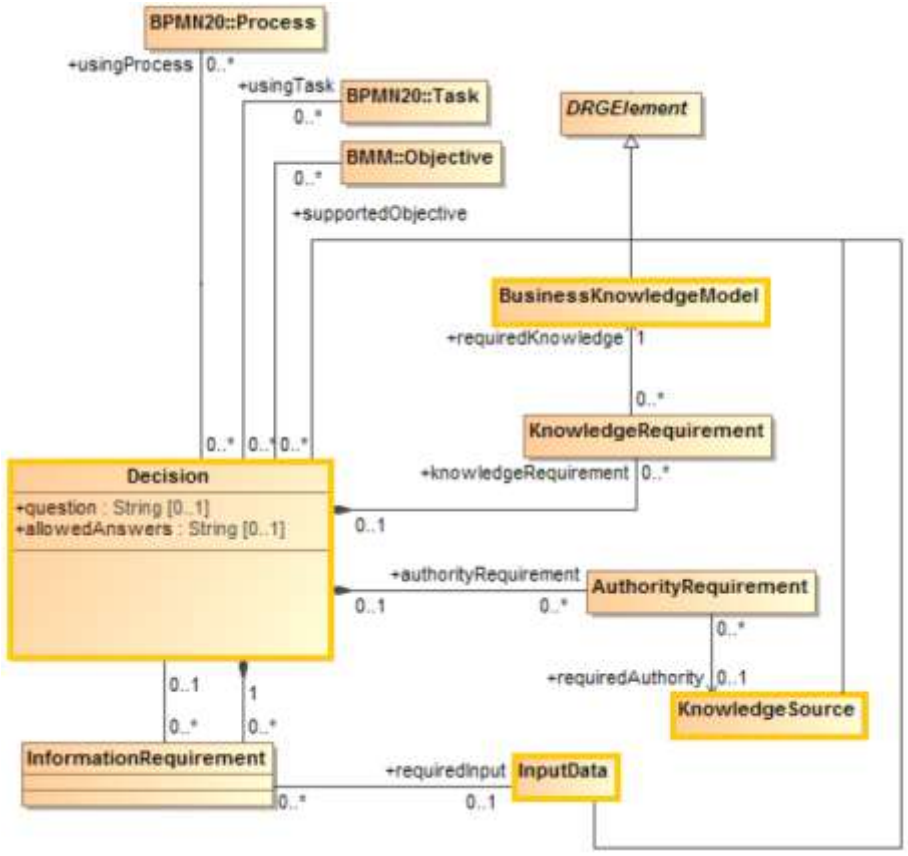


Figure 3: Extrait du métamodèle de DMN de l'OMG (diagramme de classes UML)

Sous une apparente simplicité, le Decision Requirements Diagram peut devenir rapidement complexe (si la prise de décision est elle-même complexe).

Dans l'étude de cas qui va illustrer cet article, il s'agit, dans un établissement public d'enseignement supérieur, de décider de la recevabilité (ou non) d'un vacataire lors du processus de recrutement, selon 11 critères différents.

Le DRD représenté par la Figure 4 reste encore simple. Le choix a été fait de représenter chaque critère comme une donnée d'entrée. Cela permet de voir au premier coup d'œil que les critères sont complètement différents, selon qu'il s'agit d'un poste de chargé d'enseignement ou d'un poste d'agent temporaire.

Un autre choix de modélisation aurait pu être de représenter uniquement le vacataire comme seule donnée d'entrée, les 11 critères étant alors considérés comme ses attributs.

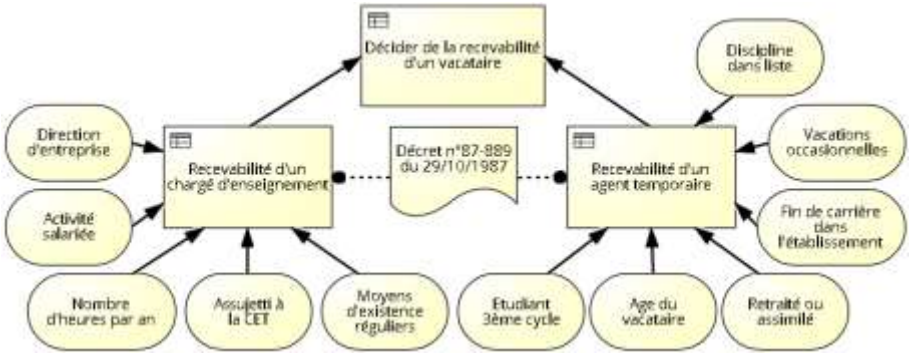


Figure 4: Diagramme (DRD) "Décider de la recevabilité d'un vacataire"

On voit aussi sur cet exemple que la décision principale « Décider de la recevabilité d'un vacataire », représentée par convention tout en haut du DRD, est constituée de deux décisions secondaires « Recevabilité d'un chargé d'enseignement » OU « Recevabilité d'un agent temporaire ».

Le Décret n°87-889 est considéré comme Source de connaissance pour chacune des décisions secondaires. Les règles métier, qu'il convient d'appliquer pour décider de la recevabilité d'un vacataire, proviennent de ce document de référence.

Par contre, cette étude de cas n'utilise pas de Modèle de connaissance métier, car la logique de décision est ici très spécifique et ne pourrait pas être réutilisée ailleurs. Un Modèle de connaissance métier s'utile comme une fonction, que l'on appelle avec des paramètres spécifiques à chaque contexte. Son objectif premier est la réutilisation.

2.2.2. Tables de décision

La notation graphique - relativement simple - du DRD ne permet pas de représenter la logique de décision elle-même. Celle-ci est représentée généralement (surtout lorsque les critères sont relativement nombreux) par une table de décision [Vanthienen et Dries, 1994].

Le formalisme choisi dans la spécification DMN pour représenter les tables de décision est celui du CODASYL, qui existe depuis quelques décennies [Codasyl, 1982]. Le Tableau 1 représente par exemple la logique pour décider de la recevabilité d'un chargé d'enseignement, en fonction de 5 critères en entrée (4 booléens et 1 nombre).

La logique pour décider de la recevabilité d'un agent temporaire est représentée par une table de décision similaire, avec des critères en entrée différents, bien sûr. Quant à la décision principale de recevabilité d'un vacataire, il s'agit d'une table très simple représentant un OU logique des deux décisions secondaires. Ces deux tables ne sont pas reproduites dans cet article, pour des contraintes de place.

Tableau 1: Table de décision "Recevabilité d'un chargé d'enseignement"

Recevabilité d'un chargé d'enseignement						
U	Entrées					Sortie
	Direction entreprise	Activité salariée	Nombre d'heures par an	Assujetti à la CET	Moyens existence réguliers	Recevabilité d'un chargé enseignement
	Booléen	Booléen	Nombre	Booléen	Booléen	Booléen
1	= vrai	-	-	-	-	vrai
2	= faux	= vrai	≥ 900	-	-	vrai
3	= faux	= vrai	< 900	-	-	faux
4	= faux	= faux	-	= vrai	-	vrai
5	= faux	= faux	-	= faux	= vrai	vrai
6	= faux	= faux	-	= faux	= faux	faux

La lettre « U » (comme Unique) positionnée en haut et à gauche de cette table de décision signifie qu'il ne doit pas de recouvrement entre chacune des 6 règles. Unique est la politique de succès (*hit policy*) qui est recommandée : pour chaque jeu de données d'entrée, une seule règle doit s'appliquer.

La formalisation des règles métier est une spécialité à part entière [Ross, 2013]. Les outils spécialisés dans la notation DMN - une vingtaine sont déjà disponibles [OpenRules, 2016] - permettent de vérifier automatiquement la complétude et la cohérence des tables de décision, ce qui est un gage de qualité.

2.2.3. Langage FEEL

FEEL (*Friendly Enough Expression Language*) est un langage déclaratif proposé dans la spécification DMN qui permet de formaliser la logique de décision sous la forme de code informatique, indépendamment de toute plate-forme [Silver, 2016]. Exemple : `if Vacataire.DirectionEntreprise = true then true`

FEEL n'est toutefois pas suffisant pour représenter tous les éléments de DMN (diagrammes et tables de décision), comme cela sera montré plus loin.

2.2.4. Complémentarité de DMN avec BPMN

La notation DMN peut s'utiliser seule ou bien en complément de la notation BPMN [Debevoise et Taylor, 2014]. Chaque Décision peut être utilisée par un processus ou une tâche, comme indiqué sur l'extrait du métamodèle (Figure 3).Nous avons montré dans un précédent article que la notation BPMN se révélait souvent mal adaptée pour représenter des prises de décisions complexes [Biard et al., 2015].

Dans un diagramme BPMN, il suffit alors de remplacer une multitude de branchements en cascade par une seule tâche de type *Business Rules*, symbolisée par une petite table de décision en haut à gauche, pour faire le lien avec un diagramme DMN. La bonne pratique est de nommer à l'identique la tâche du diagramme BPMN et la décision principale du diagramme DMN.

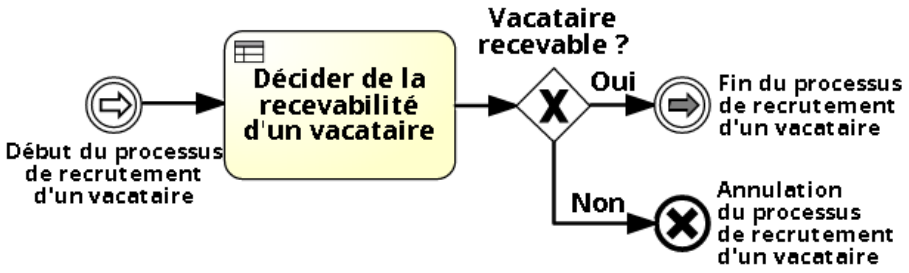


Figure 5: Extrait du diagramme BPMN Processus de recrutement d'un vacataire

Généralement, une tâche de type Business Rules est suivie par un branchement du type OU exclusif, comme dans l'exemple ci-dessus, mais pas toujours. En effet, la notation DMN peut servir à modéliser un calcul complexe, comme un pourcentage de remise en fonction de la fidélité d'un client ou le taux d'un emprunt en fonction des risques présentés par l'emprunteur : dans ce cas, la tâche retourne le résultat du calcul uniquement, qui sera transmis à la tâche suivante.

3. MDA (Model Driven Architecture)

3.1 Trois modèles CIM, PIM & PSM

La *Model Driven Architecture* (MDA) ou l'Architecture Dirigée par les Modèles est un concept proposé par l'OMG également [OMG, 2014]. Ce concept est indépendant de tout langage ou notation. La MDA applique notamment le principe de Séparations des Préoccupations [Dijkstra, 1974].

La MDA propose entre autres trois modèles à différents niveaux, qui semblent particulièrement bien adaptés pour comprendre, voire comparer, les deux solutions proposées pour automatiser les prises de décisions modélisées avec la notation DMN. Le Tableau 2 qui rassemble ces 3 modèles va nous servir de grille de lecture.

Tableau 2: Les trois modèles MDA : CIM, PIM & PSM

Trois modèles MDA	Description
CIM (Computation Independent Model)	Modèle de représentation du métier, indépendant de toute considération informatique
PIM (Platform Independent Model)	Modèle de conception pour l'informatique, indépendant de la plate-forme d'exécution
PSM (Platform Specific Model)	Modèle de conception pour l'informatique, spécifique à la plate-forme d'exécution

3.2. Projection de DMN sur les modèles de la MDA

Si l'on projette les éléments principaux de DMN présentés précédemment sur les trois modèles MDA, on obtient assez logiquement le Tableau 3 dans lequel la plate-forme cible choisie est Drools [Red Hat, 2017], le plus connu des BRMS (Business Rules Management System), qui contient entre autres un moteur de règles. Ce tableau est une première approche théorique. En effet, les expérimentations présentées dans les chapitres suivants démontreront que ce Tableau 3 devra être complété.

Tableau 3: Projection des principaux éléments de DMN sur les modèles MDA

Trois modèles MDA	Principaux éléments de DMN en théorie
CIM (Computation Independent Model)	DRD (Decision Requirements Diagram) + Tables de décision
PIM (Platform Independent Model)	FEEL (Friendly Enough Expression Language)
PSM (Platform Specific Model)	DRL (Drools Rule Language)

4. Démonstrateur pour l'automatisation des prises de décision

4.1. Présentation du démonstrateur

Notre démonstrateur est constitué essentiellement de :

- Signavio Decision Manager version 10.11.0 [Signavio, 2016] pour la modélisation en DMN des prises de décision (diagrammes et tables de décision) ; cet outil fonctionne entièrement en ligne en mode SaaS et ne requiert donc qu'un navigateur web ; il est de bonne facture et d'accès gratuit pour le monde académique. Enfin, il est complètement intégré avec Process Editor pour la modélisation BPMN ;
- Drools (version 7.0.0.Beta6) pour l'automatisation des prises de décisions ; il s'agit du BRMS de référence, qui existe en version gratuite ou payante (sous le nom de Red Hat JBoss BRMS), proposé seul ou intégré dans des logiciels plus complets ; Drools est utilisé ici sous la forme d'un *plug-in* pour Eclipse.

L'environnement de développement intégré Eclipse (version neon.2) a donc été utilisé pour faire fonctionner Drools. D'autres applications comme Git pour la gestion des versions et Maven pour la gestion des dépendances s'avèrent également utiles.

4.2. Première solution spécifique : du CIM au PSM (sans PIM)

4.2.1. Principe de la première solution spécifique

Depuis plusieurs mois, Signavio Decision Manager permet de générer directement un fichier au format DRL (Drools Rule Language), c'est-à-dire du code informatique spécifique à une plate-forme (PSM), à partir d'un diagramme DMN et des tables de décision (CIM) qui lui sont associés.

Le fichier DRL utilise, en plus des bibliothèques standards de Java, une bibliothèque de fonctions spécifiques à Signavio, fournie sous la forme d'un fichier « DMN Formulae Java8-1.0-SNAPSHOT.jar ».

Tableau 4 : Projection des éléments de la première solution sur les modèles MDA

Trois modèles MDA	Signavio Decision Manager (solution 1)
CIM (Computation Independent Model)	DRD (Decision Requirements Diagram) + Tables de décision
PIM (Platform Independent Model)	✘
PSM (Platform Specific Model)	DRL (Drools Rule Language) + DMN Formulae Java8-1.0-SNAPSHOT.jar

4.2.2 Changement de paradigme : la programmation déclarative

Même si cette solution n'utilise pas le modèle PIM de niveau intermédiaire, il est intéressant de s'y attarder, car le langage DRL ne propose pas moins qu'un changement de paradigme, la programmation déclarative [Van Roy et Haridi, 2004], par rapport à un langage informatique traditionnel (programmation impérative).

En effet, dans le langage DRL, pas de « if…then…else » imbriqués qui rendent le développement, le test et la maintenance compliqués avec les autres langages, mais uniquement un « when…else » pour chaque règle métier, que l'on peut facilement ajouter, modifier ou supprimer en toute indépendance des autres règles.

Un autre avantage majeur est que l'ordre des règles métier n'est pas important (si la politique de succès est de type Unique). Même la boucle « for each » pour prendre en compte tous les objets concernés par les règles métier n'est plus utile en programmation déclarative.

Algorithme 1: Programmation impérative avec "if...then…else"

```
for each objet in mesDonnees do
   if une première condition est détectée
      then une première action est faite
      else
         if une deuxième condition est détectée
            then une deuxième action est faite
            else une troisième action est faite
         endif
   endif
endfor
```

Algorithme 2: Programmation déclarative avec "when...then"

```
rule "une"
   when une première condition est détectée
   then une première action est faite
end
```

```
rule "deux"
   when une deuxième condition est détectée
   then une deuxième action est faite
end
```

```
rule "trois"
   when une troisième condition est détectée
   then une troisième action est faite
end
```

Les vraies règles métier de notre étude de cas s'avèrent assez verbeuses et peu lisibles, comme c'est souvent le cas du code informatique généré automatiquement. L'exemple ci-après (Algorithme 3) a été simplifié (suppression de préfixes), afin que les lignes aient une longueur raisonnable dans cet article.

Dans la section *when*, les Données d'Entrée sont évaluées en lignes 3, 4 et 5. Dans la section *then*, si les critères de décision sont respectés, la Décision RecevabiliteChargeEnseignement devient vraie en ligne 8. A noter que cette Décision est insérée dans la base de faits en ligne 9, car il s'agit d'une sous-décision pour la Décision principale DeciderRecevabilitéVacataire et sa valeur sera utilisée ultérieurement dans une autre règle.

Algorithme 3: Exemple de code DRL simplifié pour la règle n°2 du Tableau 1

```
1: rule "recevabiliteChargeEnseignement_rule_2"
2:   when
3:     eval(nullSafeEval(equals(getDirectionEntreprise(),false)))
4:     eval(nullSafeEval(equals(getActiviteSalariee(),true)))
5:     eval(nullSafeEval(greaterThanOrEqualTo(getNombreHeuresParAn(),
                                        BigDecimal.valueOf(900.0))))
6:   then
7:     RecevabiliteChargeEnseignement $recevabiliteChargeEnseignement
                             = new RecevabiliteChargeEnseignement();
8:     $recevabiliteChargeEnseignement.
                            setRecevabiliteChargeEnseignement(true);
9:     insert($recevabiliteChargeEnseignement);
10:end
```

4.2.3 Déclaration simplifiée des Données d'entrée, sans caractères accentués

Le challenge était de pouvoir utiliser les fichiers au format DRL, tel quel, sans aucune modification, afin qu'un changement de règle puisse être répercuté facilement. Seuls les caractères accentués de la langue française posent problème, car ils sont interdits dans la plupart des langages de programmation et en l'occurrence supprimés lors de la génération des fichiers au format DRL, rendant la compréhension des noms des éléments compliquée. Un diagramme et ses tables de décision modifiés, sans aucun caractère accentué, ont donc été nécessaires. Le langage DRL permet ensuite une déclaration des Données d'entrée très simple, grâce à la directive *declare*.

Algorithme 4: Déclaration des Données d'entrée simplifiée avec le langage DRL

```
1: declare Input
2:    directionEntreprise : Boolean
3:    activiteSalariee: Boolean
4:    nombreHeuresParAn: BigDecimal
5:    assujettiCet : Boolean
6:    moyensExistenceReguliers: Boolean
7: end
```

4.2.4 Application des règles métier pour un vacataire

Afin d'appliquer toutes les règles métier générées automatiquement (ligne 09), il convient d'écrire un programme en langage Java, qui va d'abord créer (ligne 02) puis initialiser (lignes 03 à 07) un objet de type vacataire, puis l'insérer (ligne 08) parmi les faits de la session Drools. Cette partie du programme est relativement simple et lisible. Ce n'est pas le cas du code qui suit (lignes 10 à 13), nécessaire à la récupération du résultat (la décision *output*). Ces quatre lignes de codes ont nécessité beaucoup d'essais, notamment à cause du typage spécifique des objets.

Algorithme 5: Programme Java d'application des règles, première solution (extrait)

```
01: FactType vacataireType =
       kieBase.getFactType("com.signavio.droolsexport.mon_modele_dmn",
          "Input");
02: Object thierry = vacataireType.newInstance();
03: vacataireType.set(thierry, "directionEntreprise", true);
04: vacataireType.set(thierry, "activiteSalariee", false);
05: vacataireType.set(thierry, "nombreHeuresParAn",
       new BigDecimal("800"));
06: vacataireType.set(thierry, "assujettiCet", true);
07: vacataireType.set(thierry, "moyensExistenceReguliers", true);
08: kieSession.insert(thierry);
09: kieSession.fireAllRules();

10: FactType outputType =
       kieBase.getFactType("com.signavio.droolsexport.mon_modele_dmn",
          "DeciderRecevabiliteVacataire_Output");
11: Collection<?> outputObjects = kieSession.getObjects(new
       ClassObjectFilter(outputType.getFactClass()));
12: Object outputObject = outputObjects.iterator().next();
13: boolean output = (boolean) outputType.get(outputObject,
       "deciderRecevabiliteVacataire");
14: System.out.println("Recevabilité du vacataire = " + output);
```

Quant à l'affichage du résultat (la décision *output*), celui-ci s'avère peu spectaculaire dans notre étude de cas, puisqu'il s'agit d'une simple variable booléenne qui représente la recevabilité du vacataire lors de son recrutement : *true* ou *false* :

```
Recevabilité du vacataire : true
```

4.3. Deuxième solution générique : du CIM au PIM (sans PSM)

4.3.1 Principe de la deuxième solution générique

Depuis quelques semaines, Signavio Decision Manager permet également de générer un fichier XML au format DMN version 1.1 (appelé « DMN 1.1 XML »), toujours à partir d'un diagramme DMN et des tables de décision (CIM) qui lui sont associés. « DMN 1.1 XML » est un format d'interchange défini dans la spécification DMN. Un schéma XSD est fourni par l'OMG, permettant de valider le fichier XML généré. Il s'agit donc d'un PIM, car il est indépendant de toute plate-forme.

Le but originel de ce format d'interchange est de pouvoir échanger des modèles de prise de décision entre des outils différents. Ce fichier XML dit sérialisé définit tous les éléments (y compris des métadonnées) nécessaires à la représentation du diagramme et des tables de décision. Ce format d'interchange est également capable de contenir du langage FEEL (qui lui n'est pas suffisant pour tout définir).

La prochaine version 7 de Drools étant capable d'interpréter directement le format « DMN 1.1 XML » (comme quelques autres outils spécialisés dans la modélisation et l'exécution des décisions), la génération de code spécifique à une plate-forme (PSM) devient alors superflue.

Tableau 5: Projection des éléments de la deuxième solution sur les modèles MDA

Trois modèles MDA	Signavio Decision Manager (solution 2)
CIM (Computation Independent Model)	DRD (Decision Requirements Diagram) + Tables de décision
PIM (Platform Independent Model)	DMN 1.1 XML (Modèle interchangeable) contenant éventuellement du langage FEEL
PSM (Platform Specific Model)	✘

4.3.2 Application des règles métier pour un vacataire

Le fichier « DMN 1.1 XML » généré par Signavio Decision Manager n'a pas fonctionné tel quel avec Drools. Quelques ajustements ont été nécessaires, essentiellement l'ajout du namespace « signavio » comme préfixe des types de variables créées avec cet outil : `<variable typeRef="signavio:monType"` (tandis que les types de variables génériques sont déjà définis dans le namespace « feel » : `<typeRef>feel:boolean</typeRef>`). Le préfixe adéquat sera sans doute ajouté dans d'une prochaine version de l'outil.

Afin d'appliquer les règles métier générées automatiquement, il convient également d'écrire un petit programme en langage Java. Alors qu'il faut ajouter au préalable deux lignes de code pour prendre le fichier « DMN 1.1 XML », la création et l'initialisation du contexte sont très similaires à celui d'un objet. Par contre, la méthode de récupération du résultat (toujours aussi peu spectaculaire) s'avère particulièrement concise, voire élégante, par rapport à celle de la première solution.

Algorithme 6: Programme Java d'application des règles, deuxième solution (extrait)

```
01: DMNRuntime runtime = DMNRuntimeUtil.createRuntime(
       "13-Decider-recevabilite-vacataire-DMN.dmn", this.getClass() );
02: DMNModel dmnModel = runtime.getModel(
       "http://www.signavio.com/dmn/1.1/diagram",
          "13-Decider-recevabilite-vacataire-DMN" );

03: DMNContext context = DMNFactory.newContext();
04: context.set( "directionEntreprise", true);
05: context.set( "activiteSalariee", false);
06: context.set( "nombreHeuresParAn", 800);
07: context.set( "assujettiCet", true);
08: context.set( "moyensExistenceReguliers", true);

09: DMNResult dmnResult = runtime.evaluateAll( dmnModel, context );
10: DMNContext result = dmnResult.getContext();
11: System.out.println( "Recevabilité du vacataire : " + result.get(
       "deciderRecevabiliteVacataire" ) );
```

4.3. Comparaison des deux solutions

La deuxième solution générique a sans aucune hésitation notre préférence. Elle est mieux alignée sur les modèles MDA que la première solution spécifique car cette deuxième solution est indépendante de toute plate-forme. Elle est très intéressante si la plate-forme cible - le moteur de règles (BRMS) - est capable d'interpréter directement des fichiers au format « DMN 1.1 XML » (ce qui reste encore assez rare).

En fait, la première solution a permis de mettre en valeur la deuxième ! La première solution a également permis de découvrir le changement important de paradigme, la programmation déclarative, qui est également utilisée de manière implicite dans la seconde solution.

5. Conclusion et perspectives

Malgré son aspect apparemment simpliste, nous avons démontré dans cet article que les prises de décision modélisées en notation DMN (diagramme et tables de décision) peuvent être automatisées. Les différents modèles de la MDA nous ont permis de porter un regard d'architecte sur les solutions techniques proposées, afin de les comprendre puis de les mettre en œuvre.

Cette mise en œuvre nécessite aussi des compétences de développeur informatique. Toutefois, avec seulement quelques lignes de code Java (de niveau avancé), il est possible d'appliquer des règles métier complexes, qui ont été définies en amont par des analystes métier. Si le jeu de données d'entrée (les critères de décision) est complet dès le départ, il est possible de modifier les règles sans changer le code qui les applique.

Tandis que l'automatisation des règles métier existe depuis une quinzaine d'années [Taylor, 2007], le fait de pouvoir désormais s'appuyer en amont sur la notation standard DMN est un progrès indéniable.

Après que les processus métier aient été extraits des applications au début des années 2000, nous pouvons présager que la prochaine étape sera l'extraction des règles métier dans les années qui viennent. L'existence de la notation standard DMN pour la modélisation des prises de décisions, et la possibilité récente d'automatiser ces prises de décisions constituent de sérieux atouts pour la réalisation de ce présage.

6. Bibliographie et références

Biard, T., Le Mauff, A., Bigand, M., Bourey, J.-P. (2015). Separation of Decision Modeling from Business Process Modeling Using New « Decision Model and Notation » (DMN) for Automating Operational Decision-Making. In L. M. Camarinha-Matos, F. Bénaben, & W. Picard (Éd.), *Risks and Resilience of Collaborative Networks* (Vol. 463, p. 489-496). Springer International Publishing.

Codasyl. (1982). A modern appraisal of decision tables. ACM.

Debevoise, T., Taylor, J. (2014). *The MicroGuide to Process and Decision Modeling in BPMN/DMN*. ACR.

Dijkstra, E. W. (1974). On the role of scientific thought.

Linehan, M., de Sainte Marie, C. (2011). The Relationship of Decision Model and Notation (DMN) to SBVR and BPMN. http://www.brcommunity.com/b597.php

OMG. (2013). Business Process Model and Notation (BPMN). http://www.omg.org/spec/BPMN/

OMG. (2014). MDA (Model Driven Architecture) Specifications. http://www.omg.org/mda/specs.htm

OMG. (2015a). Business Motivation Model (BMM). http://www.omg.org/spec/BMM/

OMG. (2015b). Semantics of Business Vocabulary and Rules (SBVR). http://www.omg.org/spec/SBVR/

OMG. (2015c). Unified Modeling Language (UML). http://www.omg.org/spec/UML/

OMG. (2016). Decision Model and Notation (DMN). http://www.omg.org/spec/DMN/

OpenRules. (2016). Decision Model and Notation (DMN) Supporting Tools. http://openjvm.jvmhost.net/DMNtools/

Pitschke, J. (2014). Mastering Business Modeling – Applying Business Architecture and Standards successfully. http://www.enterprise-design.eu/files/images/download-praesentaionen/BPMEurope2014_JPitschke.pdf

Red Hat. (2017). Drools, Business Rules Management System. http://www.drools.org

Ross, R. G. (2013). *Business rule concepts: getting to the point of knowledge* (4th Ed). Business Rule Solutions.

Signavio. (2016). Decision Manager. http://www.signavio.com/products/decision-manager/

Silver, B. (2016). *DMN method and style*. Cody-Cassidy Press.

Taylor, J. (2007). *Smart (enough) systems: how to deliver competitive advantage by automating the decisions hidden in your business*. Prentice Hall.

Van Roy, P., Haridi, S. (2004). *Concepts, techniques, and models of computer programming*. MIT Press.

Vanthienen, J., Dries, E. (1994). Decision Tables: Refining the Concept and a Proposed Standard. *ResearchGate*.

Index des auteurs

Index des auteurs

Programme de la conférence

Conférences invitées

Session 1 - Sémantique des données et connaissances

Session 2 - Modèles : concepts et ingénierie

Session 3 - Filtrage d'informations

Session 4 - Processus : concepts et ingénierie

Session 5 - Patrons de conception

Session 6 - Analyse de l'information dans les réseaux sociaux

Session 7 - Gestion de données complexes

Session 8 - Ingénierie des méthodes

Ateliers

Open et/ou Linked Data dans les systèmes d'information

Description

Les systèmes d'information (SI) exploitent différents types de données entre autres des données "liées" (Linked Data) et/ou "ouvertes" (Open Data). Cependant, l'exploitation de ce type de données dans les SI soulève des questions que l'atelier propose de discuter.

Parmi les plus importantes, et au-delà du passage à l'échelle, nous trouvons l'hétérogénéité, la sélection, le traitement, la fusion et la validation des données extraites depuis les nombreuses sources disponibles. L'impact de ces données dans les SI est d'autant plus important si l'on ajoute des problématiques décisionnelles, de mobilité, de contextes fortement contraints, d'accessibilité, de contextualisation, d'évanescence, etc.

Afin de dresser un panorama des problématiques centrées sur les Linked Data et/ou les Open Data, nous proposons dans le cadre d'INFORSID 2017 un atelier durant lequel les problématiques émergentes et/ou transverses sur l'intégration des Linked Data et/ou Open Data dans les SI pourront être partagées et discutées. Pour balayer le processus depuis la génération d'Open/Linked Data jusqu'aux consommateurs de ces données, nous encourageons les regards croisés entre acteurs de la recherche, acteurs institutionnels publics et acteurs privés dans ce domaine. Ainsi, l'atelier a pour objectif de faire se rencontrer ces différentes sensibilités.

Programme de l'atelier

- Ouverture de l'atelier
- Regards croisés sur la thématique des Open Data & des Linked Data
  - Introduction générale : Cassia Trojahn - IRIT - 15 minutes
  - Institutionnel : Fabien Moguen - Chef de projet OpenDataLab - 15 minutes
  - Entreprise : Guy Pascal - Oracle - 15 minutes
- Présentations longues (20 minutes chacune)
  - Les données de qualité logicielle, vecteur d'amélioration des logiciels de la gestion informatique, Guillaume Kerrien, Wahiba Bahsoun, Célia Nouguier - CGI - IRIT
  - Une approche originale de groupement de résultats SPARQL, Sonia Djebali, Thomas Raimbault, Pôle Universitaire Léonard de Vinci - Digital group, ESILV, France
- Présentations courtes (10 minutes chacune)
  - Projet ModRef : comparaison du langage CIDOC-CRM avec Nakala et DBPEDIA, Pascaline Kenfack-Tchienehom
  - LinkedMDR: un modèle sémantique de représentation de corpus de documents multimédia, Nathalie Charbel

  - OpenData : dans la soute d'un producteur de données ouvertes, Sandrine Mathon, Directeur de projet OpenData, Toulouse Metropole

- Clôture de l'atelier

Si le temps le permet, nous pourrons accepter des présentations courtes (10 minutes) de dernière minute. Cependant pour faciliter l'organisation de l'atelier, si vous souhaitez proposer une proposition courte, merci de prendre contact au plus vite avec les organisateurs.

Comité d'organisation

- Max Chevalier (Université de Toulouse - Paul Sabatier, IRIT, Toulouse) : Max.Chevalier@irit.fr
- Sébastien Laborie (Université de Pau et des Pays de l'Adour - LIUPPA, Bayonne) : Sebastien.Laborie@iutbayonne.univ-pau.fr
- Cassia Trojahn (Université de Toulouse - Jean Jaurès, IRIT, Toulouse) : Cassia.Trojahn@irit.fr
- Antoine Zimmermann (école des mines de Saint-étienne, LHC, Saint-étienne) : Antoine.Zimmermann@emse.fr

Date : Mardi 30 Mai 14h00-17h00

Lieu : Manufacture des Tabacs, Toulouse

Contact : Max.Chevalier@irit.fr

Site Web : http://slaborie.perso.univ-pau.fr/images/Conferences/INFORSID17-atelier/

De la surveillance à la gestion de crise : prise en compte des alertes

Description

Lors de l'édition 2016 du Congrès INFORSID, nous avons animé un atelier dédié à la diffusion des alertes et aux systèmes d'aides à la décision. De nombreux travaux se concentrent sur les systèmes de gestion de crise. Lors de cette nouvelle édition, nous avons souhaité nous confronter aux questions liées aux changements de paradigmes entre un mode d'usage normal et un mode de gestion de crise : comment s'opère ce changement d'usage ? Par un mécanisme de remontée d'information ou d'alerte, par une décision et un changement de type de gouvernance ? Qu'est-ce que cela implique dans le suivi de l'information, dans le système d'information, dans ses interfaces, dans le processus de décision et de suivi ?

Pour nous aider à répondre à ces questions, nous nous sommes focalisés sur les axes suivants :

- Gestion des activités de contrôle et de surveillance
- Gestion de crise
- Gestion des alertes et des évènements anormaux
- Mode de gouvernance et workflow
- Dynamique des SI
- Reconfiguration dynamique d'environnements complexes
- Adaptation de salle de supervision à un usage dans des domaines particuliers comme : la sécurité des infrastructures, les villes connectées, la cybersécurité, etc.

Mots-clés : SI, alerte, gestion de crise, ...

Exposés

- Florence Sèdes (introduction) "SI, alarmes et alertes : de la qualité des données à la confiance, une approche pluridisciplinaire et soci(ét)ale"
- Invité : Alexandre CABROL PERALES, responsable du laboratoire d'innovation du centre de cybersécurité de Sopra Steria à Colomiers (31), France.
  Sopra Steria apporte aux entreprises et administrations dans le monde numérique des capacités d'excellence en cybersécurité et confiance numérique pour protéger leurs informations sensibles et accélérer leur développement numérique. Avec plus de 700 experts en sécurité mobilisés en Europe et ses centres de cybersécurité, Sopra Steria dispose d'une force de frappe humaine, technologique et industrielle de premier ordre.

Invités / table-ronde

- Agence Eau Adour Garonne (31)
- Y. Bardie, EDEC-MRM-SI, Thérèse Libourel, ESPACE-DEV, Montpellier, "Vers un modèle de sûreté systémique dans les organisations de pharmacovigilance"
- Maude Arru, Elsa Nègre, Camille Rosenthal-Sabroux, LAMSADE, Université Paris-Dauphine, "Vers une modélisation des comportements en situation de crise"
- Slimane Hammoudi, Nicolas Gutowski et Olivier Camp, ESEO, Angers, "Mobilité et prédiction : un atout pour la gestion de crise"
- Antoine Humeau, ONFCS, Détection automatique en temps proche du réel de signaux anormaux de mortalité dans la faune sauvage
- Wafa Abdelghani, Corinne Amel Zayani, Florence Sedès, Ikram Amous, IRIT, Toulouse et Université de Sfax, Sfax, Tunisie, "Vers une gestion des alertes via le crowdsoucing et la confiance dans l'IoT"
- Table-ronde / conclusion et future works

Date : Mardi 30 Mai 14h00-17h00

Lieu : Manufacture des Tabacs, Toulouse

Contact : Florence.Sedes@irit.fr

Enseignement des SI

Description

L'atelier "Enseignement des SI" a pour objectif de mieux cerner comment mettre en œuvre des modalités pédagogiques non-traditionnelles en co-concevant une séquence pédagogique pluri-disciplinaire. Il vise à identifier les modalités pédagogiques non traditionnelles favorisant la pluridisciplinarité et à étudier l'intégration des modalités pédagogiques non traditionnelles dans un enseignement des SI pluridisciplinaire.

L'atelier présentera rapidement un état des pratiques pédagogiques envisageables en se basant sur les expériences des participants (par exemple, les approches agiles, par le jeu ou tangible). Dans un deuxième temps, une étude de cas basée sur la mise en œuvre d'un module de "Analyse, conception et développement d'applications" serait proposée.

Animatrices

- Gaëlle Blanco-Lainé, Université Grenoble Alpes
- Sophie Dupuy-Chessa, Université Grenoble Alpes, LIG
- Jannik Laval, Université Lumière Lyon 2, DISP Lab

Date : Mardi 30 Mai 14h00-17h00

Lieu : Manufacture des Tabacs, Plateau pédagogie Active de la BU

Contact : Sophie.Dupuy@imag.fr

VADOR : Valorisation et Analyse des Données de la Recherche

Description

Cet atelier s'insère dans une dynamique émergente sur l'analyse de données de la recherche (données numériques produites par les chercheurs, mémoires, articles scientifiques, actes de colloque, thèses, etc.) et veut faire la promotion de la recherche francophone.

Il permettra d'aborder des thématiques variées faisant cohabiter des disciplines différentes autour de la problématique de l'analyse et de la valorisation des données de la recherche, d'un point de vue théorique ou pratique.

Programme prévisionnel

- 14h - 15h : Conférence invitée de Chérifa Boukacem-Zeghmouri, MCF HDR, Lyon 1, Elico, Urfist de Lyon.
  Données de la recherche et publication scientifique : à la lumière des valeurs et des régulations de l'Open Science
  Dans les discours performatifs et opératoires autour de l'Open Science, les données de la recherche et la publication scientifique sont le plus souvent présentées séparément, animées par des objectifs spécifiques. La réalité est toute autre et mérite que l'on s'y attarde. Les liens entre données de la recherche et publications scientifiques se font par des régulations socio-techniques (TDM, web sémantique) et socio-économiques (modèles d'affaires, stratégies d'acteurs, tarifications) qui, ensemble, tissent l'Openness d'une science qui se pratique déjà par certaines communautés. Autour de ces liens, se développent des stratégies, des services, des produits et des métiers qui donnent à voir les perspectives et les opportunités qui se profilent pour de nouvelles pratiques d'élaboration scientifiques. Adossée au cadre théorique des industries culturelles et créatives, cette intervention vise à rendre compte de la manière avec laquelle les liens "données" et "publications", qui se construisent sous nos yeux, incarnent une réalité concrète de l'Open Science qui révèle beaucoup des nouvelles valeurs qui y ont cours.
- 15h - 15H40 : 2 articles (20 minutes par article incluant questions)
- 16h - 17h20 : 4 articles
- 17h20 - 17h30 : conclusion et perspectives

Date : Mercredi 31 Mai 14h00-17h00

Lieu : Manufacture des Tabacs, Toulouse

Contact : vador@sciencesconf.org

Site Web : https://vador.sciencesconf.org/

Data Visualisation dans les Systèmes d'Information

Description

Cet atelier vise à faire rencontrer les chercheurs, industriels et étudiants autour du thème de la visualisation de données. Chercheurs et professionnels sont invités à venir alimenter une réflexion commune sur les nouvelles technologies en Data Visualisation et leurs apports aux systèmes d'information ou bien en tant qu'outil d'aide à la décision. Des questions sur la qualité des données ainsi que les droits d'accès et la sécurité des données seront également explorées pendant l'atelier.

Programme prévisionnel

- 14h - 14h30 : Data visualisation: what is at stake?, Christophe Bontemps, Toulouse School of Economics (INRA)
- 14h30 - 15h30 : La visualisation de données pour le web : le nec plus ultra ?, Alain Ottenheimer de la société Datasens
- 15h30 - 15h50 : pause-café
- 15h50 - 16h10 : Présentation d'articles sélectionnés
- 16h10 - 17h : Etude de cas sur la visualisation de données.
  Intervenant : Olivier Dao Hodac de la société FlightWatching.
  Titre de l'intervention : Visualisations et analyses: cas d'Utilisation dans l'aéronautique
- 17h - 17h30 : Conclusion et perspectives

Date : Mercredi 31 Mai 14h00-17h00

Lieu : Manufacture des Tabacs, Toulouse

Contact : eliana.raad@univ-pau.fr

Systèmes d'information et Démocratie

Description

Cet atelier se propose d'interroger l'influence que les systèmes d'information et de décision peuvent exercer sur la démocratie dans les organisations, et, plus largement, dans la société. L'atelier prendra la forme de courtes présentations suivies de tables rondes favorisant la participation de tous et la discussion autour des sujets présentés.

Dans ce contexte, on abordera les questions suivantes : quels sont les impacts des systèmes d'information pour la démocratie dans l'entreprise ? Les systèmes d'information renforcent-ils les rapports de force établis ? Technologies de l'information : la démocratie peut-elle y survivre ?

Programme prévisionnel

La première table ronde consacrée à la **conception et à l'usage des systèmes d'information dans l'entreprise** :

- Raphaëlle Bour (Informatique), Université Toulouse 1 Capitole : *Les processus de conception des systèmes d'information sont-ils des processus démocratiques ?*
- Gabriel Collettis (Economie), Université Toulouse 1 Capitole : *Domination de l'information financière et conception de l'entreprise dans un capitalisme financiarisé*
- Salvatore Maugeri (Sociologie), IUT de Chartres (sous réserve)
- Maryse Salles (Informatique), Université Toulouse 1 Capitole : *Les systèmes d'information d'aide à la décision dans les entreprises favorisent-ils la démocratie ?*

La seconde table ronde traitera du thème **Technologies de l'information, Big Data et algorithmes** :

- Céline Castets-Renard (Droit), Université Toulouse 1 Capitole : *Droit, éthique et décision algorithmique*
- Nikos Smyrnaios (Information et Communication), Université de Toulouse - Paul Sabatier : *L'infomédiation au cœur des systèmes qui régissent notre vie numérique*
- Jean-Sébastien Vayre (Sociologie), Université de Toulouse - Jean Jaurès : *Les systèmes de traitement de mégadonnées favorisent-ils la démocratie ?*

Date : Jeudi 1 Juin 14h00-17h00

Lieu : Manufacture des Tabacs, Toulouse

Contact : maryse.salles@ut-capitole.fr

Combiner des données d'observation satellitaire avec d'autres sources pour l'aide à la décision et l'intelligence spatiale

Description

Dans le cadre du programme européen Copernicus mais aussi, de dispositifs nationaux (Equipex GEOSUD, Pôle THEIA) les données d'observation de la terre deviennent de plus en plus accessibles, que ce soit au niveau de la communauté scientifique, des entreprises ou auprès du grand public. En parallèle, l'ouverture massive de données numériques sous différentes formes (spatialisées ou non) et relatives à différents domaines et leur mise en relation avec les données satellitaires, ouvrent de nouveaux champs d'applications.

Cet atelier vise à rassembler les chercheurs, PMEs, industriels et étudiants autour des enjeux liés au croisement des données d'observation de la Terre avec d'autres sources de données en particulier ouvertes (données INSEE, réseaux sociaux, données citoyennes ou volontaires, capteurs terrestres, données Google StreetView, publications scientifiques, ...) pour le développement d'outils l'aide à la décision.

Programme prévisionnel

- W. Heintz, J. Mothe, N. Neptune, J.-B. Puel. Combiner des données d'observation satellitaire avec d'autres sources pour l'aide à la décision et l'intelligence spatiale
- F. Renard, L. Alonso, W. Bechkit, L. Ponsar. La combinaison de l'image satellitaire avec les données citoyennes pour la mesure de l'îlot de chaleur urbain, Présentation du projet Cit'Air
- Zoé, Mathieu, François et Cédric de la Geek'O Team. Projet FLL Animal Allies : Play4Wild, Une façon moderne de lutter contre le braconnage des éléphants
- J.B. Puel. Observatoire Multimédia du Paysage, images aériennes et images au sol pour analyser et enseigner les dynamiques paysagères
- D. Sheeren, M. Fauvel, A. Vialatte, D. Dallery, V. Thierion, W. Heint. L'usage des images très haute résolution spatiale pour une meilleure prédiction des services écosystémique

Date : Jeudi 1 Juin 14h00-17h00

Lieu : Manufacture des Tabacs, Toulouse

Contact : josiane.mothe@irit.fr

Sécurité des systèmes d'information : technologies et personnes, 2e édition

Description

La sécurité des systèmes est au plus haut niveau d'intérêt des chercheurs, des entreprises et de la population. Cependant, en ce qui concerne la sécurité des systèmes d'information (SI), il n'existe ni processus souverain, ni modèle consensuel. En effet, un SI peut être confronté à des menaces non seulement à l'extérieur, mais aussi à l'intérieur même de l'organisation qu'il supporte. Ces menaces peuvent être de accidentelles et non malicieuses à intentionnelles et malicieuses. Dans ce contexte, les employés sont susceptibles de constituer une menace pour la sécurité du SI dès lors qu'ils en sont considérés comme des composants. En tant que tels ils constituent, tout comme les artefacts technologiques, un point d'entrée qu'il ne faut pas négliger. L'objectif de cet atelier est d'être un espace d'échange entre chercheurs et professionnels afin d'alimenter une réflexion commune sur la sécurité des SI, d'un point de vue non seulement technologique, mais aussi humain.

Programme prévisionnel

Première partie - présentation des articles sélectionnés (1h15)

- Ronan Champagnat et Mourad Rabah
  Laboratoire L3i, Université de La Rochelle
  Apport de l'utilisation de la fouille de processus pour améliorer la sécurité des SI (Fouille de processus, politique d'accès aux données)

- Dhavy Gantsou
  LAMIH, CNRS UMR 8201, Université de Valenciennes et du Hainaut Cambrésis
  Cyberattaques internes : quelles approches technologiques pour la prévention et la défense ? (Cyber sécurité, menaces internes, défense active, défense proactive, attaques avancées)

- Wilson Goudalo, Christophe Kolski et Frédéric Vanderhaegen
  Université de Valenciennes et du Hainaut Cambrésis et Research and Innovation Department, Advanced Business Engineering
  Démarche d'ingénierie de la sécurité dans le management de projet : activités de sécurité et relations entre acteurs (Sécurité, Acteurs, Rôles, Management de Projet, Risques de Sécurité)

- Raogo Kabore, Yvon Kermarrec et Philippe Lenca
  Lab-STICC, IMT Atlantique
  Revue des systèmes de détection d'anomalies dans les réseaux SCADA(Détection d'anomalies, SCADA, IDS, intrusion, ICS, systèmes de contrôle industriels)

- Zakariya Kamagate, Yvon Kermarrec et Jacques Simonin
  Lab-STICC, IMT Atla

Architecture logicielle et détection d'anomalies (Ingénierie Dirigée par les Modèles (IDM), Sécurité, Ingénierie Logicielle, Détection d'anomalies)

- David Sheeren, Mathieu Fauvel, Aude Vialatte, Donatien Dallery, Vincent Thierion, Wilfried Heint
  Groupe de recherche en intelligence artificielle et gestion (GRIAGES), Université Catholique d'Afrique Centrale, Yaoundé, Cameroun
  GRIAGES, Université Catholique d'Afrique Centrale, Yaoundé, Cameroun
  Modèle d'évaluation de l'impact d'un système de management de la sécurité de l'information sur la performance

Deuxième partie - table ronde (1h15)

Échanges autour de questions émergentes en sécurité des systèmes d'information

Date : Vendredi 2 Juin 10h30-13h00

Lieu : Manufacture des Tabacs, Toulouse

Contact : pierre-emmanuel.arduin@dauphine.fr

Résumé

Ce document contient les actes du trente-cinquième congrès INFORSID (INFormatique des ORganisations et Systèmes d'Information de Décision) qui s'est déroulé à Toulouse du 30 mai au 2 juin 2017. Le processus de sélection des articles publiés a été organisé à deux niveaux avec un Conseil du Comité de Programme (CoP) additionnel au Comité de Programme habituel (CP). Les membres du CoP ont organisé une méta-évaluation d'un pool d'articles qui leur ont été affectés. La méta-évaluation a consisté à organiser les discussions entre les membres du CP relecteurs de chaque article afin de résoudre les conflits d'évaluation et d'aboutir, dans la mesure du possible, à un consensus. Les membres du CoP ont rédigé, à la fin du cycle de discussions, une brève évaluation de synthèse pour chacun des articles de leur pool d'articles.