



**Actes du Congrès**

# **INFORSID 2022**

INFormatique des Organisations et Systèmes  
d'Information et de Décision

40<sup>ème</sup> édition

**Dijon, 31 mai au 3 juin 2022**

Présidente du Comité de Programme : Nathalie Vallès-Parlangeau

Présidente du Comité d'Organisation : Lylia Abrouk

<https://inforsid2022.sciencesconf.org/>



# Préface

## *Y a-t-il un bel âge ?*

Sans doute non, mais il y a des « *chiffres ronds* », qui toujours nous font penser qu'une étape est franchie, laissant derrière elle un sentiment de nostalgie et entre-ouvrant un monde de tous les possibles. C'est ce monde des possibles que nous allons explorer tous ensemble, membres de cette belle communauté de 40 années.

Et oui, après deux années « *à distance* », enfin ensemble pour une édition Dijonnaise qui s'est tant fait attendre.

INFORSID (INFormatique des ORganisations et Systèmes d'Information et de Décision) s'est bâtie aux quatre coins de l'Hexagone (mais pas que...), autour de rencontres pluridisciplinaires au cœur de l'informatique : le Système d'Information (SI). Depuis sa 1<sup>ère</sup> édition en 1983, il constitue un moment d'échange privilégié d'une communauté de chercheurs et praticiens pour dessiner le présent et le futur du SI. C'est grâce à cette synergie que naissent des collaborations, des projets, des thèses...

## *Des temps pour se souvenir...*

- Jean-Pierre Giraudin (Université Grenoble Alpes) fera la charnière entre hier et demain en ouvrant le congrès avec *Préhistoire et histoire d'INFORSID : une communauté francophone de recherche en SI*
- Guillaume Cabanac (Université Toulouse 3) et Cécile Favre (Université Lyon 2) proposeront une présentation à deux voix pour porter un regard scientométrique avec *INFORSID : 40 années d'une aventure scientifique*

## *Accueillante et ouverte....*

En 2022, trois conférenciers invités nous font le plaisir et l'honneur de nous rejoindre :

- François Le Gac et Eric Tassone (AIRBUS) viendront proposer une alternative aux GAFAM avec *le CLOUD souverain une perspective française et européenne*,
- Guillaume Cabanac, « Inforsidien » qui fait partie des 10 personnes qui ont le plus marqué la science en 2021 (Nature's) viendra nous parler de *Pollution de la littérature scientifique*.

Nous ouvrons cette édition sur une journée d'ateliers de travail, qui se poursuit le second jour pour certains. Je suis particulièrement fière que ces ateliers soient empreints d'une réelle pluridisciplinarité, chère à INFORSID. Les ateliers *SI pour les humanités*

*numériques, Traces numériques pour la mobilité et Défi pour l'IOT* rapprochent les chercheurs en informatique des chercheurs en SHS (histoires, art, philosophie...), en géographie, géomatique ou encore en agriculture et environnement, pour ne citer que ceux-là. Un atelier transversal sur *l'enseignement des SI* permet d'envisager l'avenir de la discipline dans un champ informatique parfois tiré par la technique.

Nous les remercions tous de leur contribution à l'ouverture et l'animation d'INFORSID 2022.

### *Tournée vers notre avenir : nos plus jeunes...*

Nos jeunes chercheurs, futurs fers de lance de notre communauté seront tour à tour, Dragon et Chevalier dans des joutes verbales qui nous permettront de découvrir les thématiques qui nous préoccupent aujourd'hui. Qu'ils continuent ainsi de partager leur passion avec nous pendant de nombreuses années ! Merci à Elena Kornyshova pour l'organisation de cette session.

Il est important de féliciter aussi ! Cette année nous avons sept doctorants en lice pour le prix de thèse. L'heureux lauréat est Paul Boniol, qui a soutenu sa thèse à l'Université de Paris en collaboration EDF R&D, sous la direction de Themis Palpanas, Emmanuel Remy, et Mohammed Meftah.

### *Pour une science en mouvement...*

Depuis quelques années, nous sommes fiers d'une session internationale qui permet d'élargir nos horizons et partager avec la communauté nationale les travaux les plus récents, reconnus dans de très bonnes conférences ou revues internationales. Neuf articles sont présentés cette année.

Nous avons cette année reçu vingt-huit soumissions d'articles originaux. Les auteurs ont contribué de tous les horizons francophones (Algérie, Belgique, Congo, France, Suisse, Tunisie). Quatorze articles ont été acceptés sous différents formats : onze articles longs, un article court, deux articles de démonstrations.

Dans un premier temps, chaque article soumis a été évalué par trois membres du Comité de Programme. Une réunion à distance du Conseil du Comité de Programme a permis de sélectionner les articles. Quelques articles ont nécessité une meta-revue et des « berger(e)s » ont suivi l'évolution de ces articles. Sur les quatre articles dans ce cas, trois ont été retenus. Le processus de sélection a été supporté par l'outil EasyChair.

Les vingt-trois articles (nationaux et internationaux) couvrent un large panel des problématiques autour des SI répartis dans le programme en plusieurs sessions : conception, modèles et méthodes, processus métiers, transformation digitale, sécurité, santé et mobilité.

### *Enfin...*

Cette édition ne se tiendrait pas sans le concours des membres du Comité de programme international (France, Belgique, Canada, Luxembourg, Suisse) et aux membres du Conseil du Comité de Programme. Leur évaluation précieuse, efficace et constructive a garanti la qualité de cette 40<sup>ème</sup> édition. Je les en remercie.

Au terme de cette édition, il convient de remercier l'ensemble de notre communauté :

- les auteurs pour leurs contributions de qualité
- les membres du Conseil du Comité de Programme pour leurs conseils pendant le processus de sélection des articles
- les porteurs des ateliers qui œuvrent pour ouvrir la communauté à des problématiques grâce à des échanges fructueux et originaux
- les trois conférenciers invités et les intervenants « spécial anniversaire » pour avoir animé le congrès et partagé avec nous leur vision du SI passé ou « avenir » et posé un regard critique sur la recherche
- les présidents de session pour leur participation active
- et enfin, les nombreux participants au Congrès, qui sont l'âme et le corps de notre communauté.

Je remercie Franck Ravat, président de l'association INFORSID, mais aussi et surtout les membres du bureau de m'avoir fait confiance pour l'organisation scientifique du congrès. Une mention spéciale à Cécile Favre pour son accompagnement sans faille dans la construction de ces actes.

Je remercie vivement l'ensemble du comité d'organisation Dijonnais, et tout particulièrement sa présidente Lylia Abrouk. Ils ont réalisé un travail de tous les instants avant et surtout pendant, pour une édition « relevée » mais qui « ne nous monte pas au nez » !

Deiz-ha-bloaz laouen, Zorionak zuri, Gaujós aniversari, joyox aniversèro... bref très joyeux anniversaire à tous,

Nathalie VALLES-PARLANGÉAU  
Présidente du Comité de Programme INFORSID 2022



## Comités

Le comité de la 40ème 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.

### Comité de Programme

Raphaëlle Bour, Université Toulouse 1 Capitole, France  
Khalid Benali, Université Nancy Lorraine, France  
Mireille Blay-Fornarino, Université Nice Sophia Antipolis, France  
Laurence Capus, Université Laval, Québec, Québec  
Faiza Ghozzi, ISIMS, Sfax, Tunisie  
Marianne Huchard, Université Montpellier, France  
Régine Laleau, Université Paris-Est Créteil, France  
Jannick Laval, Université Lyon 2, France  
Eric Leclercq, Université de Bourgogne, France  
Imen Megdiche, Université Toulouse 3, France  
André Miralles, Université Montpellier, France  
Elsa Nègre, Université Paris-Dauphine, France  
Raquel A. Oliveira, Université Grenoble, France  
Thomas Polacsek, Onera, Toulouse, France  
Christophe Ponsard, Université de Namur, Belgique  
Jolita Ralyte, Université de Genève, Suisse  
Philippe Ramadour, Aix-Marseille Université, France  
Philippe Roose, Université de Pau et des Pays de l'Adour, France  
Marinette Savonnet, Université de Bourgogne, France  
Carine Souveyet, Université Paris1, France  
Dalila Tamzalit, Université de Nantes, France  
Hervé Verjus, Université Savoie Mont-Blanc, France  
Robert Viseur, Ecocentric, Belgique

### Conseil du Comité

Julien Aligon, Université Toulouse 1 Capitole, France  
Eric Andonoff, Université Toulouse 1 Capitole, France  
Rebecca Deneckere, Université Paris1, France  
Cyril Faucher, Université de La Rochelle, France  
Cécile Favre, Université Lyon 2, France  
Agnès Front, Université Grenoble, France  
Sébastien Laborie, Université de Pau et des Pays de l'Adour, France

**Présidente :** Nathalie Vallès-Parlangeau, IRIT-UT1, Université Toulouse 1 Capitole

## **Comité d'organisation**

Claire Bourgeois-Republique, LIB - Université de Bourgogne  
Hamza Chergui, LIB - Université de Bourgogne  
Hocine Cherifi, LIB - Université de Bourgogne  
Christophe Cruz, LIB - Université de Bourgogne  
Alexis Guyot, LIB - Université de Bourgogne  
Marinette Savonnet, LIB – Université de Bourgogne

**Présidente** : Lylia Abrouk, LIB - Université de Bourgogne

## **Porteurs d'Ateliers**

Sandro Bimonte, TSCF, INRAE  
Cyril Faucher, La Rochelle Université, L3i, France  
Alexandre Journaux, GenPhySE, INRAE  
Stéphane Lamassé, LAMOP–Université de Paris 1  
Eric Maldonado, Direction Systèmes d'Information-Unité d'Appui, INRAE  
Káthia Marçal de Oliveira, Univ. Polytechnique Hauts-de-France  
Cédric du Mouza, CEDRIC-CNAM  
Manuel Munier, Université de Pau et des Pays de l'Adour  
Vincent Negre, LEPSE, INRAE  
Matthieu Noucher, CNRS, UMR Passages, Bordeaux  
Christian Sallaberry, LIUPPA, Pau  
Chantal Soulé-Dupuy, IRIT-UT1, UT1 Capitole  
Nathalie Vallès-Parlangeau, IRIT-UT1, UT1 Capitole  
Didier Vye, La Rochelle Université, UMR LIENSs



## Table des Matières

### Conférences invitées

Le Cloud Souverain <i>François Le Gac, Eric Tassone</i> .....	1
Pollution de la littérature scientifique : détection participative d'expressions torturées révélatrices d'articles frauduleux <i>Guillaume Cabanac</i> .....	3

### Mobilité

Mesure de similarité pour les trajectoires sémantiques : prise en compte de trois niveaux de granularité <i>Cécile Cayère, Christian Sallaberry, Cyril Faucher, Marie-Noëlle Bessagnet, Philippe Roose, Maxime Masson</i> .....	5
Innovation collaborative pour la mobilité des seniors - Une démarche expérimentale de conception de services de covoiturage solidaire <i>Christine Verdier</i> .....	21

### Conception

Lambda+ Architecture et vérification de la conservation des propriétés avec la théorie des catégories <i>Annabelle Gillet, Éric Leclercq, Nadine Cullot</i> .....	41
Les exigences pour un choix d'architecture dans le cadre d'une migration dans les nuages <i>Antoine Aubé, Thomas Polacsek, Clément Duffau</i> .....	43
Profilage de services pour la gestion de l'énergie dans les architectures orientées services <i>Jorge Andrés Larracochea, Philippe Roose, Sergio Ilarri, Yudith Cardinale, Sébastien Laborie</i> .....	59

### Graphe

Vers une approche efficace de gestion d'évolution des données graphes <i>Landy Andriamampianina, Franck Ravat, Jiefu Song, Nathalie Vallès-Parlangeau</i> ...	61
Towards a Graph-Oriented Perspective for Querying Music Scores <i>Philippe Rigaux, Virginie Thion</i> .....	63
Détection de signaux faibles : une méthode basée sur les graphlets <i>Hiba Abou Jamra, Marinette Savonnet, Éric Leclercq</i> .....	79
Spygraph : un robot d'exploration léger dédié à l'analyse de graphes d'hyperliens <i>Robert Viseur</i> .....	81

## Processus métiers

Une méthode pour la modélisation de la variabilité des indicateurs de performance des processus intégrée dans des modèles de processus variables  
*Diego Diaz, Mario Cortes-Cornax, Agnès Front, Cyril Labbé, David Faure* ..... 85

Traitement des événements complexes pour une gestion proactive des instances d'un processus métier  
*Abir Ismaïli-Alaoui, Khalid Benali, Karim Baïna* ..... 87

Analyse des Approches de Fouille d'Intentions : un Cadre de Comparaison  
*Rebecca Deneckère, Elena Kornyshova, Charlotte Hug* ..... 103

## Transformation digitale

Anti-patrons d'alignement métier des SI - Proposition de classification issue d'une expérience professionnelle  
*Jean-Philippe Gouigoux, Dalila Tamzalit* ..... 105

Vers un modèle d'équilibre de maturité numérique pour les organisations publiques  
*Mateja Nerima, Jolita Ralyté* ..... 107

## Sécurité

Méta-modèle des concepts et processus d'analyse des risques selon les normes de cybersécurité  
*Christophe Ponsard, Valery Ramon, Mounir Touzani* ..... 109

Pirate ta fac ! Ludification de séances de cours sur la sécurité des systèmes d'information  
*Pierre-Emmanuel Arduin, Benjamin Costé* ..... 125

Détection de fraude financière dans un système de transactions interbancaires.  
*Hamza Chergui, Lylia Abrouk, Nadine Cullot, Nicolas Cabioch* ..... 141

Construction d'une ontologie dans le domaine financier pour la détection de fraudes  
*Benjamin Auger, Hamza Chergui, Yara Chehade, Jana El Kadri, Lylia Abrouk, Nicolas Cabioch* ..... 157

## Modèles et méthodes

Évaluation de la valeur des données - Modèle et méthode  
*Jacky Akoka, Isabelle Comyn-Wattiau* ..... 163

Application de l'Ingénierie des Exigences basée sur les Modèles dans Trois Grands Projets Collaboratifs Européens : Un Rapport d'Expérience  
*Andrey Sadovykh, Hugo Bruneliere, Dragos Truscan* ..... 179

Perception des Méthodes Agiles par les Développeurs Aujourd'hui  
*Florian Gauthier, Rébecca Deneckère* ..... 181

## Santé

Extraire et organiser des connaissances sur les pathogènes - Projet EPICURE

*Leïla Renard, Thérèse Libourel, Catherine Moulia, Laurent Gavotte*..... 197

ParkinsonCom : Outil d'Aide à la Communication pour Personnes atteintes de la Maladie de Parkinson

*Káthia Marçal de Oliveira, Nejmeddine Allouche, Véronique Delcroix,*

*Yohan Guerrier, Christophe Kolski, Sophie Lepreux, Philippe Pudlo,*

*Yosra Rekik, Elise Batselé, Mathilde Boutiflat, Mark Freyens, Hélène Geurts,*

*Romina Rinaldi, Loïc Dehon, Nicolas Jura* ..... 213



---

## Le Cloud Souverain

**François Le Gac et Eric Tassone**

*AIRBUS, France*

---

*RESUME.*

*Airbus prépare actuellement un appel d'offre pour sélectionner le fournisseur de cloud européen souverain qui hébergera les backbones critiques PLM et ERP. 4 dimensions sont considérées en terme de souveraineté: les données, les opérations, les logiciels et les matériels. L' appel d'offre se concentrera sur les deux premières dimensions essentiellement.*

*Airbus recherche un catalogue de services exhaustif et évolutif couvrant les services managés: IaaS, CaaS et PaaS avec des niveaux de services très élevés. Airbus veut s' affranchir des risques légaux d'extraterritorialité et du risque de dépendance à un fournisseur non européen afin de garantir une continuité de ses opérations. 17 catégories de besoins sont exprimées dans cet appel d'offre qu'Airbus compte lancer sur le marché, fin 2022-début 2023. Une harmonisation des besoins des grands industriels sera poussée par Airbus durant l'année 2022.*

---

### **Eléments biographiques**

François Le Gac

François Le Gac a construit sa carrière au sein d'Airbus sur les domaines d'infrastructure : Réseaux, Back Office Opération. Il a rejoint l'entité Digitale en 2017 en tant que responsable de la plateforme IOT. Il est aujourd'hui en charge des plateformes Cloud Publiques et la stratégie multi cloud pour le groupe Airbus depuis 2019.

Eric Tassone

Eric Tassone a bâti sa carrière à Airbus et a évolué sur les domaines de l'infrastructure : Réseaux, Opération applicative, Entreprise Etendue. Il a participé à la mise en place de la plateforme DevOps en 2017 et a rejoint l'équipe Cloud depuis Juin 2021. Il est en charge du projet EU Cloud Souverain.

---

## **Pollution de la littérature scientifique : détection participative d'expressions torturées révélatrices d'articles frauduleux**

**Guillaume Cabanac**

*IRIT UMR 5505 CNRS, Université Paul Sabatier, Toulouse, France*  
[guillaume.cabanac@irit.fr](mailto:guillaume.cabanac@irit.fr)

---

*RESUME.*

*Nous avons découvert des milliers de publications non fiables dans les catalogues des maisons d'édition de premier plan : Elsevier, Springer et Wiley, notamment. Publiés et souvent vendus, ces pseudo-articles générés par ordinateur ou assemblés par des paper mills<sup>123</sup> tels des patchworks sont trahis par la présence d'« expressions torturées » dénuées de sens. Cet exposé présentera la plateforme 'Problematic Paper Screener'<sup>4</sup> pour identifier cette pollution affectant la littérature scientifique. Avec d'autres détectives scientifiques et lanceurs d'alerte, nous l'employons pour ré-évaluer les 6 000 articles identifiés à ce jour et les signaler sur la plateforme d'évaluation post publication PubPeer pour les faire rétracter. Cette initiative bénévole de fact-checking participatif détecte de nouvelles phrases torturées et méconduites qui sont intégrées au système développé, conduisant à étendre le détecteur par effet boule de neige. Les 6 000 articles problématiques sont parus majoritairement depuis 2014 et font l'objet de 40 000 citations en tout. Des centaines d'entre eux sont abusivement cités, sans logique apparente, indice d'une manipulation visant à augmenter le nombre de citations de certains fraudeurs.*

*Cet exposé reprendra notre article paru dans le Bulletin of the Atomic Scientists créé en 1945 par les scientifiques du projet Manhattan, traitant de « la sécurité mondiale et les questions de politique publique, en particulier celles liées aux dangers posés par les armes nucléaires et autres armes de destruction massive » (Wikipedia<sup>5</sup>).*

---

<sup>1</sup> <https://forbetterscience.com/2020/01/24/the-full-service-paper-mill-and-its-chinese-customers/>

<sup>2</sup> <https://www.nature.com/articles/d41586-021-00733-5>

<sup>3</sup> <https://forbetterscience.com/2021/05/26/the-chinese-paper-mill-industry-interview-with-smut-clyde-and-tiger-bb8/>

<sup>4</sup> <https://www.irit.fr/~Guillaume.Cabanac/problematic-paper-screener>

<sup>5</sup> [https://fr.wikipedia.org/wiki/Bulletin\\_of\\_the\\_Atomic\\_Scientists](https://fr.wikipedia.org/wiki/Bulletin_of_the_Atomic_Scientists)

*Cabanac, G., Labbé, C., & Magazinov, A. (2022). "Bosom peril" is not "breast cancer": How weird computer-generated phrases help researchers find scientific publishing fraud. Bulletin of the Atomic Scientists. <https://thebulletin.org/2022/01/bosom-peril-is-not-breast-cancer-how-weird-computer-generated-phrases-help-researchers-find-scientific-publishing-fraud/>*

---

### **Eléments biographiques**

Guillaume Cabanac est maître de conférences habilité à diriger des recherches (HDR) en informatique à l'Université Toulouse III – Paul Sabatier. Il est membre de l'Institut de Recherche en Informatique de Toulouse (IRIT UMR 5505 CNRS) et siège au Comité national du CNRS en qualité de membre nommé du Conseil scientifique de l'Institut des sciences humaines et sociales (InSHS). Ses travaux interdisciplinaires contribuent à l'analyse de la littérature scientifique, notamment au sein de l'ERC Synergy 'Nanobubbles' questionnant le processus d'auto-correction en science. Il développe la plateforme 'Problematic Paper Screener'<sup>6</sup> qui signale des milliers d'articles non fiables, pourtant publiés et souvent vendus par les maisons d'édition de premier plan. Cette recherche a été distinguée dans le "Nature's 10"<sup>7</sup> présentant « dix personnes qui ont aidé à façonner la science en 2021 » selon la revue Nature.

---

<sup>6</sup> <https://www.irit.fr/~Guillaume.Cabanac/problematic-paper-screener>

<sup>7</sup> <https://www.nature.com/immersive/d41586-021-03621-0>



---

## Mesure de similarité pour les trajectoires sémantiques : prise en compte de trois niveaux de granularité

Cécile Cayère<sup>1</sup>, Christian Sallaberry<sup>2</sup>, Cyril Faucher<sup>1</sup>,  
Marie-Noëlle Bessagnet<sup>2</sup>, Philippe Roose<sup>2</sup>, Maxime Masson<sup>2</sup>

1. La Rochelle Université,  
23 Avenue Albert Einstein,  
17000 La Rochelle, France  
cecile.cayere1@univ-lr.fr, cyril.faucher@univ-lr.fr

2. Université de Pau et des Pays de l'Adour,  
Avenue de l'Université,  
64000 Pau, France  
christian.sallaberry@univ-pau.fr, marie-noelle.bessagnet@univ-pau.fr, philippe.roose@iutbayonne.univ-pau.fr, maxime.masson@univ-pau.fr

---

**RÉSUMÉ.** Ce papier s'inscrit dans le cadre du projet DA3T en collaboration avec des géographes. L'objectif est de proposer des méthodes et outils ayant pour objectif de traiter des traces de mobilité. Ce travail est expérimenté dans le domaine du tourisme en vue d'améliorer l'analyse de mobilité de touristes et par conséquent l'aménagement et la valorisation du territoire. Dans le cadre de la conception d'un module de calcul de similarité entre des trajectoires sémantiques, nous présentons une nouvelle mesure de similarité en vue de comparer deux déplacements selon trois dimensions (c.-à-d. spatiale, temporelle et thématique) et leurs trois niveaux de granularité (c.-à-d. micro, méso et macro).

**ABSTRACT.** This paper is part of the DA3T project in collaboration with geographers. The objective is to propose methods and tools to process mobility traces in order to improve their analysis and consequently touristic territory planning and valorization. As part of the design of a module for computing similarity between semantic trajectories, we present a new similarity measure for comparing two trips on three dimensions (i.e. spatial, temporal and thematic) and three granularity levels (i.e. micro, meso and macro).

**MOTS-CLÉS :** mesure de similarité ; trajectoire sémantique ; trace de mobilité

**KEYWORDS:** similarity measure ; semantic trajectory ; mobility track

---

## 1. Introduction

La traçabilité de la mobilité humaine est un phénomène qui prend beaucoup d'ampleur de par l'évolution des technologies GPS et l'augmentation des déplacements humains. Dans le projet régional Nouvelle-Aquitaine DA3T (c.-à-d. Dispositif d'Analyse des Traces numériques pour la valorisation des Territoires Touristiques), nous exploitons les traces laissées par des touristes afin d'aider les décideurs locaux dans la gestion et l'aménagement des territoires touristiques. Il s'agit d'un projet pluridisciplinaire réunissant informaticiens et géographes dans l'objectif de produire des outils et des méthodes d'analyse de traces de mobilité.

Dans le cadre de ce projet, nous avons développé une application mobile, nommée Geoluciole, permettant de capturer les déplacements de touristes volontaires, auxquels nous avons fait passer des entretiens semi-directifs afin d'obtenir plus d'informations sur ces déplacements (p. ex. activités touristiques pratiquées). Nous avons conçu un modèle générique de description de trajectoires sémantiques et une plateforme modulaire permettant la conception et l'exécution de chaînes de traitement dédiées aux traces de mobilité. Ainsi, cette plateforme permet de paramétrer et d'enchaîner des modules de traitement de bas niveau en vue de répondre à un questionnement de plus haut niveau sur un jeu de traces de mobilité. Nous avons expérimenté ces propositions sur différents questionnements et jeux de données dédiés au tourisme, à la migration de colonies d'oiseaux ou encore aux activités d'observation menées par des naturalistes.

Dans cet article, nous nous intéressons à un module particulier de la plateforme. Il s'agit du module de calcul de similarité de deux trajectoires sémantiques conçu pour la comparaison de déplacements touristiques. Le travail présenté ici, considère les dimensions spatiale, temporelle et thématique de deux trajectoires sémantiques pour établir leur degré de similarité. Nous nous positionnons dans les domaines informatique et géomatique pour traiter des données de capteurs enrichies. Le verrou scientifique relève de la recherche d'information géographique (RIG) : il s'agit de proposer une nouvelle métrique de comparaison de trajectoires combinant trois dimensions et trois niveaux de granularité pour chaque dimension. L'originalité tient dans l'hypothèse de travail qui consiste à observer chaque couple de trajectoires selon trois dimensions à un niveau micro, méso et macro successivement.

L'article est organisé comme suit. La section 2 présente quelques définitions relatives aux trajectoires sémantiques et illustre nos motivations grâce à un scénario utilisant le jeu de données touristiques relatif à l'été 2020 à La Rochelle. La section 3 fait l'état de l'art des mesures de similarité (et de distance) dédiées aux trajectoires sémantiques. La section 4 rappelle les besoins des géographes, détaille notre hypothèse de travail et présente notre nouvelle métrique dédiée au calcul de similarité de trajectoires sémantiques. La section 5 évalue cette mesure au travers d'une expérimentation. Pour finir, la section 6 conclut cet article et propose quelques perspectives.

## 2. Trajectoires sémantiques touristiques

L'objet central de nos recherches est la trace de mobilité touristique, elle représente le déplacement d'un objet mobile (p. ex. un touriste) à travers une suite de positions géolocalisées et horodatées. Nous construisons des trajectoires brutes à partir de ces traces selon les besoins de l'analyse. En effet, le concept de trajectoire représente la sous-partie de la trace qui a un intérêt pour une application donnée (Parent *et al.*, 2013). Dans nos travaux, les trajectoires sont construites sur des critères spatiaux et/ou temporels (p. ex. la trace d'une semaine d'un touriste pourrait résulter en un ensemble de trajectoires journalières; on s'intéresse à l'activité d'un touriste durant une journée). Les trajectoires brutes peuvent ensuite être enrichies avec des données externes et deviennent des trajectoires sémantiques.

Ces données d'enrichissement peuvent être de simples labels (p. ex. "Tour de la Lanterne") ou des objets complexes (p. ex. nom : "Tour de la Lanterne", type : "tour", localisation : [46.15579, -1.15712], etc.) et sont liées à la trajectoire : entière, à un segment ou à une position de celle-ci. Nous enrichissons les trajectoires avec des objets complexes pouvant représenter n'importe quel phénomène du monde réel, appelés aspects (Mello *et al.*, 2019). Dans notre modèle, un aspect est lié à la trajectoire par l'intermédiaire d'un ou plusieurs épisodes (c.-à-d. un intervalle temporel) qui définissent la ou les parties de la trajectoire enrichies par l'aspect. Les aspects d'un même type sont liés à une même séquence d'épisodes représentant un axe thématique particulier, appelée interprétation de la trajectoire (p. ex. séquence d'épisodes météorologiques).

Une trajectoire sémantique a une dimension temporelle (c.-à-d. les *timesteps*), une dimension spatiale (c.-à-d. les coordonnées spatiales) et un ensemble de dimensions sémantiques (c.-à-d. les interprétations).

Deux trajectoires sémantiques sont plus ou moins similaires sur une ou plusieurs de leurs dimensions. Par exemple, deux touristes peuvent suivre un même itinéraire sans pour autant pratiquer les mêmes activités ou se déplacer sur une même temporalité. Nous souhaitons comparer deux trajectoires sémantiques en tenant compte de toutes ces dimensions afin d'identifier si deux touristes ont des comportements similaires.

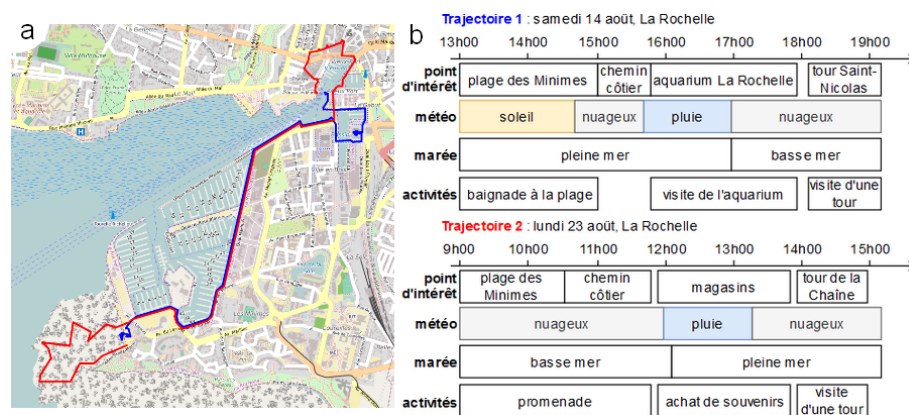


FIGURE 1. Deux trajectoires sémantiques appartenant à deux touristes différents

Prenons un exemple : comparons manuellement deux trajectoires sémantiques afin d'identifier les éléments essentiels de comparaison. La figure 1 montre deux trajectoires sémantiques construites à partir de traces collectées. La partie a de la figure illustre cartographiquement la dimension spatiale des trajectoires et la partie b met en évidence la dimension temporelle à travers un axe du temps ainsi que la dimension thématique grâce à la représentation des différentes interprétations des trajectoires. Les types d'aspects considérés ici sont les points d'intérêt, la météo, la marée et les activités touristiques (issues des entretiens).

**Dimension spatiale (c.f. figure 1, a) :** Des similitudes spatiales sont clairement visibles entre les trajectoires 1 et 2 (respectivement, bleue et rouge). Les deux se situent au centre-ville de La Rochelle et ont un point de départ et d'arrivée plus ou moins similaires (dans les mêmes zones). Les deux trajectoires comportent deux phases plutôt stationnaires ou de visite (c.-à-d. les zones de départ et d'arrivée où les positions sont plus proches les unes des autres) séparées par une phase de déplacement (c.-à-d. les longues lignes sans détour où les positions sont plus éloignées les unes des autres). De plus, il est à noter que les deux touristes traversent la ville en suivant le même chemin. Cependant, les arrêts (c.-à-d. amas de points aux mêmes endroits) que nous pouvons identifier à l'oeil nu ne sont pas les mêmes. Pour améliorer la comparaison, il faudrait pouvoir comparer les deux séquences de coordonnées géographiques. Nous pouvons déduire de toutes ces analyses que, malgré quelques légères différences, les deux trajectoires sont très similaires sur le plan spatial.

**Dimension temporelle (c.f. figure 1, b) :** Côté dimension temporelle, les deux se passent au mois d'août (c.-à-d. l'été). Ensuite, nous pouvons remarquer que l'une des trajectoires se passe le week-end (c.-à-d. un samedi) et l'autre en semaine (c.-à-d. un lundi). La durée des deux déplacements est de 7 heures environ mais ils ne se déroulent pas aux mêmes moments de la journée (l'un se déroule l'après-midi de 13h00 à 19h00, l'autre de 9h00 à 15h00). Ainsi, nous pouvons conclure que mis-à-part la saison et le mois, les trajectoires sont plutôt différentes sur le plan temporel.

**Dimension thématique (c.f. figure 1, b) :** Pour finir, concernant la dimension thématique, quatre interprétations enrichissent la trajectoire (à savoir, les points d'intérêt traversés par le touriste, la météo, la marée et les activités touristiques mentionnées dans l'entretien). Chaque épisode d'une interprétation (p. ex. "plage des Minimes") est un aspect décrit par un ensemble d'attributs non représenté ici pour ne pas surcharger la figure. La plus longue sous-séquence partagée par deux trajectoires est appelée plus longue séquence commune (Vlachos *et al.*, 2002). En considérant uniquement les points d'intérêt traversés par les touristes, la plus longue séquence commune aux deux trajectoires est ⟨plage des Minimes, chemin côtier⟩. Nous pouvons aller plus loin grâce aux types (c.-à-d. l'attribut "type") des points d'intérêt, ce qui donne la plus longue séquence commune ⟨plage, chemin, tour⟩. En nous intéressant à toutes les interprétations en même temps, nous obtenons la plus longue séquence commune suivante : ⟨plage des Minimes, (plage des Minimes, nuageux), (chemin côtier, nuageux), chemin côtier, pluie, (pleine mer, pluie), nuageux, (visite dune tour, nuageux)⟩. En observant cette séquence, on se rend compte que les trajectoires sont plutôt similaires sur le plan thématique.

La similarité (inverse de distance) entre deux trajectoires peut être évaluée grâce à une fonction de similarité (ou de distance) permettant d’attribuer un score qui varie selon leur ressemblance ou leur différence. Un score de similarité est élevé lorsque les trajectoires se ressemblent et faible lorsqu’elles diffèrent; inversement, un score de différence est élevé lorsque les trajectoires diffèrent l’une de l’autre et faible lorsqu’elles se ressemblent. De nombreuses fonctions de calcul de similarité se basent sur une ou plusieurs dimensions des trajectoires, quelques unes sont présentées dans la section suivante.

### 3. Travaux connexes

Cette section a pour but de faire le tour des mesures de similarité (et de distance) existantes permettant d’évaluer la ressemblance de deux trajectoires sur une ou plusieurs dimensions. Il existe déjà plusieurs travaux s’intéressant à comparer et classifier les mesures permettant de comparer des trajectoires (Wang *et al.*, 2013; Magdy *et al.*, 2015; Cleasby *et al.*, 2019; Su *et al.*, 2020; Tao *et al.*, 2021), cependant la dimension thématique est souvent omise, n’étant pas la dimension centrale de description du déplacement d’un objet mobile. Dans un premier temps, nous présentons les mesures de similarité spatiale. Dans un second temps, nous présentons les mesures de similarité temporelle. Dans un troisième temps, nous présentons les mesures de similarité thématique. Enfin, nous abordons le cas particulier des mesures s’intéressant aux séries temporelles qui peuvent être utilisées pour comparer les différentes dimensions des trajectoires.

La dimension spatiale d’une trajectoire GPS est une suite de coordonnées GPS, c.-à-d. une suite de paires (*longitude, latitude*) qui représente plus ou moins fidèlement l’itinéraire emprunté par l’objet mobile. Pour calculer la similarité spatiale entre deux trajectoires, nous pouvons les considérer comme des suites de points, comme des suites de segments ou nous pouvons les simplifier à leurs enveloppes englobantes. Cela revient à calculer la similarité entre des points, entre des lignes ou entre des polygones.

Pour calculer la distance entre deux points d’une trajectoire, il est possible d’utiliser la distance euclidienne (c.-à-d.  $L_2$  Norm) ou la distance de Manhattan (c.-à-d.  $L_1$  Norm). La distance euclidienne (ou ED) (Faloutsos *et al.*, 1994) peut être appliquée sur les points d’un espace euclidien à une dimension (p. ex. sur des éléments de séries temporelles) ou plusieurs dimensions (p. ex. sur des points de trajectoires). Pour mesurer la distance entre deux trajectoires dans un espace euclidien, il est possible d’utiliser la distance euclidienne entre les points correspondants des deux trajectoires (distance entre le  $i$ -ème point d’une trajectoire avec le  $i$ -ème point de l’autre trajectoire) puis d’additionner toutes les distances calculées. C’est la distance euclidienne à étapes bloquées (*lock-step euclidean distance*) (Tao *et al.*, 2021). Pour calculer la distance entre deux points GPS (sans conversion vers un espace euclidien), il existe la formule de Haversine utilisée pour la première fois en 1805 par James Andrew (An-

drew, 1805).

D'autres mesures de similarité spatiale se basent sur la division des trajectoires en segments ou en sous-trajectoires qu'elles comparent deux à deux comme la mesure SpADe (*Spatial Assembling Distance*) (Y. Chen *et al.*, 2007), la distance de Hausdorff (Alt, 2009), AMSS (*Angular metric for shape similarity*) (Nakamura *et al.*, 2013) et TRACCLUS (*TRAjectory CLUStering*) (Lee *et al.*, 2007). Nous réutilisons TRACCLUS qui compare deux segments sur trois éléments importants (c.-à-d. parallélisme, distance et angle).

Nous souhaitons comparer les polygones englobants les trajectoires. Pour cela, nous pouvons utiliser le système de raisonnement RCC-8 (*Region Connection Calculus*) qui étend les relations entre intervalles temporels d'Allen aux polygones spatiaux (Aiello, 2002)(Sallaberry, 2013) (p. ex. deux polygones sont déconnectés, se superposent, etc.) ou les 9-intersections (Egenhofer, 1997) qui décrivent les relations topologiques pouvant s'appliquer à des polygones, des lignes et des points. La mesure de similarité appliquée à la recherche d'information spatiale présentée dans Le Parc-Lacayrelle *et al.* (2007) s'appuie sur l'intersection de deux polygones pour évaluer leur similarité, avec un score nul lorsqu'il n'y a pas d'intersection. Nous réutilisons cette dernière mesure car elle permet de comparer deux trajectoires sur un gros grain de détail en utilisant leurs boîtes englobantes.

La dimension temporelle d'une trajectoire GPS est une suite de marqueurs temporels (*timestamps*). Chaque marqueur est lié à un point de la trajectoire ; le tout représente le déplacement de l'objet mobile observé. La similarité temporelle entre deux trajectoires est souvent calculée de pair avec la dimension spatiale. Cependant, nous nous intéressons dans cette section à la dimension temporelle uniquement. Pour calculer cette similarité, nous pouvons considérer les trajectoires comme des suites de marqueurs temporels ou des intervalles temporels.

Les relations d'Allen (Allen, 1983) sont un ensemble de 13 relations entre intervalles temporels (p. ex. les intervalles sont égaux, se rencontrent, etc.). Pour une paire d'intervalles donnés, ces relations renvoient des résultats booléens.

La mesure de similarité appliquée à la recherche d'information temporelle présentée dans Le Parc-Lacayrelle *et al.* (2007) s'appuie sur l'intersection entre deux intervalles temporels pour évaluer leur similarité, avec un score nul lorsqu'il n'y a pas d'intersection. Nous réutilisons cette mesure car elle permet de comparer deux trajectoires sur un gros grain de détail en utilisant les intervalles de temps englobants des trajectoires.

La dimension thématique d'une trajectoire sémantique est un ensemble d'interprétations, c.-à-d. un ensemble de séquences d'épisodes temporels liés à des aspects d'un certain type (p. ex. météo, points d'intérêt, etc.). Ainsi, chaque position correspond à un certains nombres d'épisodes appartenant à différentes interprétations dont il faut tenir compte dans le calcul de similarité. Dans les travaux connexes, afin d'évaluer la similarité thématique des trajectoires sémantiques, ces dernières sont considérées comme des suites de données d'enrichissement simples (p. ex. label) ou complexes (p. ex. aspects) (chacune correspondant à une position), des suites d'épisodes sémantiques liés à des données d'enrichissement simples ou complexes ou des labels prin-

cipaux résumant des interprétations spécifiques des trajectoires. Il existe des mesures spécifiquement destinées à comparer des trajectoires multi-aspects comme la mesure TRAFOS (Varlamis *et al.*, 2021) ou la mesure MUITAS (*MUltiple-aspect TrAjec-tory Similarity*) (May Petry *et al.*, 2019). Nous réutilisons MUITAS car elle compare des trajectoires multi-aspect telles que nos trajectoires, des seuils sont appliqués pour comparer chaque attribut de chaque aspect et des pondérations régulent l'importance de chaque type d'aspect dans le calcul du score global. De plus, contrairement à TRAFOS, MUITAS ne nécessite pas l'utilisation de toutes les trajectoires du jeu de données pour calculer la similarité entre deux trajectoires. Certaines mesures considèrent les trajectoires comme des suites d'épisodes sémantiques comme la mesure LBS-Alignment (Lu, Tseng, 2009) ou la distance d'édition enrichie (Moreau *et al.*, 2018) où toute donnée d'enrichissement comparée doit être considérée au sein d'une ontologie ou hiérarchie de concepts pour être comparée.

Les mesures LCSS/LCS (*Longest Common Subsequence*) (Vlachos *et al.*, 2002), EDR (*Edit Distance on Real sequence*) (L. Chen *et al.*, 2005), ERP (*Edit distance with Real Penalty*) (L. Chen, Ng, 2004) et DTW (*Dynamic Time Warping*) (Keogh, Ratanamahatana, 2005) sont des mesures permettant de comparer des séries temporelles qui peuvent être utilisées dans notre contexte. Elles consistent à étudier la proximité des éléments des séries, deux à deux, pour choisir les meilleures correspondances et calculer un score de similarité (ou de distance) final. Plus les éléments mis en correspondance sont éloignés, plus grande sera la pénalité ajoutée au score final. Ces mesures prennent en compte l'ordre entre les éléments mais pas l'écart temporel qui les séparent. Elles peuvent être utilisées pour calculer la similarité des différentes dimensions de la trajectoire. Par exemple, nous réutilisons DTW pour la similarité spatiale qui est très adaptée au calcul de la similarité spatiale car elle utilise directement la distance entre les points sans seuil (dont la valeur peut dépendre de la taille de la trajectoire) et EDR pour la similarité des séquences thématiques car la correspondance des aspects peut être contrôlée avec un seuil.

#### **4. Mesure de similarité DA3T dédiée aux trajectoires sémantiques**

La mesure de similarité DA3T est dédiée à la comparaison de trajectoires sémantiques de mobilité. Nous commençons par rappeler les besoins exprimés par les partenaires du projet. Nous détaillons ensuite l'hypothèse sur laquelle repose cette nouvelle métrique. Enfin, nous présentons la mesure DA3T.

##### **4.1. Rappel des besoins**

Dans le projet DA3T, les géographes veulent comparer des trajectoires deux à deux (p. ex. comparaison de trajectoires représentatives appartenant à deux catégories de visiteurs différentes, comparaison d'une trajectoire de touriste avec un parcours type de l'Office du tourisme, etc.). Ils souhaitent une mesure paramétrable (pour les dimensions spatiale, temporelle et thématique) et calculée automatiquement car ce

type de comparaison est long et, à ce jour, uniquement réalisé par des experts en géographie du tourisme.

#### 4.2. *Hypothèses de travail pour une nouvelle mesure*

Nous faisons l'hypothèse que si chaque dimension, spatiale, temporelle et thématique est observée selon trois niveaux de granularité, respectivement micro, méso et macro, nous calculons un score de similarité entre deux trajectoires avec plus de précision. Le tableau 1 illustre ces niveaux de granularité.

Concernant la **dimension spatiale**, à l'échelle micro (c.f. tableau 1, 1), nous comparons deux trajectoires point à point. À l'échelle méso (c.f. tableau 1, 2), nous comparons des sous-parties (segments) de ces trajectoires afin de découvrir si elles ont la même tendance générale. Ainsi, par exemple, pour deux trajectoires de touristes ayant empruntés la même rue sur une partie de leurs déplacements, nous pouvons détecter une forte similarité au grain méso. Enfin, à l'échelle macro (c.f. tableau 1, 3), nous comparons deux trajectoires par rapport à la taille, la forme et le chevauchement de leurs boîtes englobantes (*bounding-box*). Cela permet d'identifier si deux objets mobiles ont globalement les mêmes comportements de déplacement.

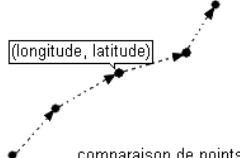
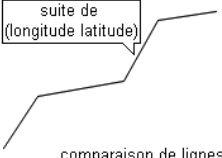
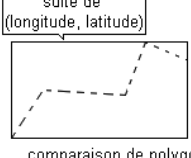
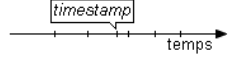

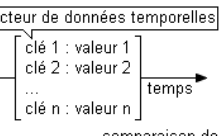
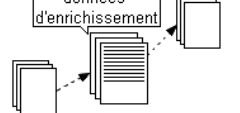
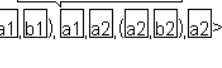
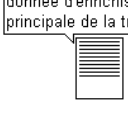
Concernant la **dimension temporelle**, à l'échelle micro (c.f. tableau 1, 4), nous comparons les *horodatages* des positions des deux trajectoires. À l'échelle méso (c.f. tableau 1, 5), comme pour le spatial, nous comparons des segments/intervalles temporels correspondants à des couples de points horodatés. Enfin, à l'échelle macro (c.f. tableau 1, 6), nous comparons le contexte temporel relatif à chaque trajectoire pour savoir si elles se déroulent durant une même année, une même saison, un même mois, un même jour de semaine, etc. Ainsi, par exemple, une trajectoire journalière se passant un mardi du mois d'août 2020 et une autre s'étalant sur un week-end du mois d'août 2021 sont en partie similaires car elles se déroulent toutes les deux au mois d'août, mais pas la même année ni le même jour de la semaine.

Enfin, concernant la **dimension thématique**, à l'échelle micro (c.f. tableau 1, 7), nous comparons les données d'enrichissement associées aux positions des trajectoires. Par exemple, prenons un point d'une trajectoire enrichi avec les données suivantes : {nom : "Tour de la Lanterne", catégorie : "tour", ...} et {description : "soleil", température : 30, ...} et un point d'une seconde trajectoire enrichi avec : {nom : "Tour de la Lanterne", catégorie : "tour", ...} et {description : "soleil", température : 11, ...} Les deux points sont similaires sur plusieurs attributs : les touristes était proche du même point d'intérêt et il faisait soleil, mais il ne faisait pas la même température. À l'échelle méso (c.f. tableau 1, 8), nous comparons des séquences de données thématiques. Par exemple, deux trajectoires enrichies avec le découpage administratif et décrites respectivement par les séquences ⟨Les Minimes, Saint-Nicolas, Les Minimes⟩ et ⟨Les Minimes, Saint-Nicolas, Centre-ville⟩, se ressemblent sur les deux premiers éléments de la séquence mais ne sont pas identiques. Pour finir, à l'échelle macro (c.f. tableau 1, 9), nous comparons les thématiques dominantes. Prenons par exemple deux trajectoires de touristes enrichies avec des données de météo ; l'une peut être résumée



par le label "nuageux", l'autre par le label "ensoleillé". Ces deux trajectoires sont donc globalement différentes.

TABLEAU 1. *Tableau récapitulatif des dimensions et niveaux de granularité d'une trajectoire sémantique*

	Micro	Méso	Macro
Spatial	1  (longitude, latitude) comparaison de points	2  suite de (longitude latitude) comparaison de lignes	3  suite de (longitude, latitude) comparaison de polygones
Temporel	4  timestamp temps comparaison de marqueurs temporels	5  paire de timestamps temps comparaison d'intervalles temporels	6  vecteur de données temporelles clé 1 : valeur 1 clé 2 : valeur 2 ... clé n : valeur n temps comparaison de vecteurs de données
Thématique	7  données d'enrichissement comparaison des données enrichissant les positions	8  séquence de données d'enrichissement $\langle (a_1, b_1), (a_1, a_2), (a_2, b_2), a_2 \rangle$ comparaison de séquences de données	9  donnée d'enrichissement principale de la trajectoire comparaison de données d'enrichissement

Ainsi, nous faisons l'hypothèse (H1) qu'introduire différents niveaux de granularité (c.-à-d. micro, méso et macro) pour chacune des dimensions (c.-à-d. spatiale, temporelle et thématique) permettra d'améliorer les mesures existantes.

#### 4.3. Mesure de similarité DA3T

Pour valider l'hypothèse (H1), nous mettons en place une formule de calcul de score de similarité qui combine des sous-scores spatial, temporel et thématique pondérés par des coefficients. Chacun de ces sous-scores est à son tour la combinaison de trois scores de différents niveaux de granularité (c.f. tableau 1). Nous expérimentons ensuite la formule sur un jeu de données issues d'une campagne d'étude de mobilité touristique dans la ville de La Rochelle. Nous comparons les résultats obtenus avec ceux issus de formules de comparaisons existantes ainsi qu'avec l'avis d'experts en géographie.

Notre mesure est définie par l'équation 1.

$$S_{\text{glb}} = \alpha_{\text{spt}} * S_{\text{spt}} + \beta_{\text{tmp}} * S_{\text{tmp}} + \gamma_{\text{thm}} * S_{\text{thm}} \quad (1)$$

Dans cette formule, chaque sous-fonction de calcul de similarité relatif à une dimension spécifique peut être détaillée en trois niveaux de granularité, telle que :

$$S_{spt} = \alpha_{spt-mic} * S_{spt-mic} + \alpha_{spt-mes} * S_{spt-mes} + \alpha_{spt-mac} * S_{spt-mac} \quad (2)$$

$$S_{tmp} = \beta_{tmp-mic} * S_{tmp-mic} + \beta_{tmp-mes} * S_{tmp-mes} + \beta_{tmp-mac} * S_{tmp-mac} \quad (3)$$

$$S_{thm} = \gamma_{thm-mic} * S_{thm-mic} + \gamma_{thm-mes} * S_{thm-mes} + \gamma_{thm-mac} * S_{thm-mac} \quad (4)$$

La somme des coefficients de pondération d'un même niveau (c.-à-d.  $\alpha_*$ ,  $\beta_*$  et  $\gamma_*$ ) est toujours égal à 1 telle que :  $\alpha_{spt} + \beta_{tmp} + \gamma_{thm} = 1$ ,  $\alpha_{spt-mic} + \alpha_{spt-mes} + \alpha_{spt-mac} = 1$ ,  $\beta_{tmp-mic} + \beta_{tmp-mes} + \beta_{tmp-mac} = 1$  et  $\gamma_{thm-mic} + \gamma_{thm-mes} + \gamma_{thm-mac} = 1$ . De plus, toute mesure de similarité  $S$  est telle que :  $0 \leq S \leq 1$ . Plus le score de similarité est proche de 1, plus les trajectoires sont similaires ; plus le score de similarité est proche de 0, plus elles sont différentes.

Dans l'équation 1, nous ré-utilisons certaines mesures de similarité existantes. Premièrement pour les mesures de la dimension spatiale (c.f. équation 2), nous utilisons les suivantes :

- $S_{spt-mic} \rightarrow$  DTW (Keogh, Ratanamahatana, 2005) et distance de Haversine (Andrew, 1805) : Pour comparer les trajectoires à l'échelle des points, nous utilisons la mesure DTW qui a pour avantage d'utiliser directement la distance spatiale entre les points pour calculer la distance total entre les trajectoires. Pour évaluer la distance entre les points, nous utilisons la distance de Haversine qui permet de mesurer la distance entre deux points GPS sur le plan terrestre. Ainsi, nous n'avons pas besoin de convertir nos données vers un espace euclidien.

- $S_{spt-mes} \rightarrow$  TRACCLUS (Lee *et al.*, 2007) : Pour comparer les trajectoires à l'échelle des segments, nous utilisons la mesure TRACCLUS car elle prend en compte différentes caractéristiques des segments (c.-à-d. leur parallélisme, leur distance et leur angle) pour les comparer et chacun des ces types de comparaison peut également être pondéré.

- $S_{spt-mac} \rightarrow$  Mesure de similarité appliquée à la RI spatiale (Le Parc-Lacayrelle *et al.*, 2007) : Pour comparer les boites englobantes des trajectoires, nous utilisons la mesure de similarité appliquée à la RI spatiale qui utilise l'intersection entre les polygones pour calculer leur distance.

Deuxièmement, pour les mesures de la dimension temporelle (c.f. équation 3), nous utilisons les suivantes :

- $S_{tmp-mic} \rightarrow$  EDR (L. Chen *et al.*, 2005) appliqué à des séries de labels : Pour comparer les trajectoires à l'échelle des *timestamps*, nous attribuons une période de la journée (p. ex. matin, après-midi, soir, etc.) à chaque *timestamp* et nous utilisons EDR pour comparer deux suites de périodes associées aux trajectoires. Nous considérons qu'il y a correspondance entre deux périodes si elles sont exactement égales.

- $S_{tmp-mes} \rightarrow$  Mesure de similarité appliquée à la RI temporelle (Le Parc-Lacayrelle *et al.*, 2007) : Pour comparer les trajectoires à l'échelle des intervalles temporels, nous ramenons les intervalles temporels des trajectoires à une échelle jour-

nalière. Nous appliquons ensuite la mesure de similarité appliquée à la RI temporelle qui utilise l'intersection entre les intervalles pour calculer leur distance.

–  $S_{tmp-mac}$  → Mesure de similarité de vecteurs de données : Pour comparer les trajectoires à l'échelle du contexte temporel, nous associons un vecteur de données temporelles à la trajectoire entière (p. ex. année : 2020, saison : automne, mois : 11, etc.) et nous comparons deux trajectoires sur leur vecteurs de données. Chaque éléments des vecteurs qui diffèrent apportent une pénalité au score.

Enfin, troisièmement, pour les mesures de la dimension thématique (c.f. équation 4), nous utilisons les suivantes :

–  $S_{thm-mic}$  → MUITAS (May Petry *et al.*, 2019) : Pour comparer les trajectoires à l'échelle des données enrichissantes les positions, nous utilisons la mesure MUITAS car elle permet de comparer des trajectoires multi-aspects en tenant compte de tous attributs de chaque aspect.

–  $S_{thm-mes}$  → EDR (L. Chen *et al.*, 2005) appliqué à des séries d'épisodes : Pour comparer les trajectoires à l'échelle des séquences d'épisodes sémantiques, nous utilisons l'attribut principal des aspects pour construire les séquences et nous exécutons EDR sur ces séquences. Lorsque nous travaillons avec des aspects lié à une ontologie, nous avons pour but d'utiliser la distance d'édition enrichie (Moreau *et al.*, 2018).

–  $S_{thm-mac}$  → LCSS (Vlachos *et al.*, 2002) : Pour comparer les trajectoires à l'échelle des thématiques générales, nous les résumons avec les valeurs majoritaires que prennent les attributs principaux de chaque type d'aspect et nous utilisons la mesure LCSS, qui est très adaptée pour comparer deux chaînes de caractères, pour comparer ces valeurs.

## 5. Expérimentation

Nous avons mis en place une expérimentation pour valider notre mesure. Elle a été conçue pour répondre à plusieurs questionnements :

1. Est-ce que les scores obtenus avec les mesures sont conformes à l'avis d'experts en géographie ?
2. Quels coefficients de pondération optimisent les résultats de la mesure ?
3. Dans le contexte de la mobilité touristique, est-il pertinent de considérer les trois dimensions des trajectoires sémantiques (c.-à-d. spatiale, temporelle et thématique) ?
4. Est-ce que notre hypothèse de départ (H1) est validée par cette expérimentation ?

### 5.1. Corpus de trajectoires

Le corpus utilisé pour cette expérimentation contient 23 paires de trajectoires journalières de touristes issues de la campagne de collecte Geoluciole. Ces paires représentent une variété de cas différents où les trajectoires peuvent être semblables

sur toutes, plusieurs, une ou aucune des dimensions présentées précédemment. Nous avons enrichi ces trajectoires avec des données de météo, de lever et de coucher de soleil issues d'OpenWeatherMap, des données concernant les points d'intérêt de La Rochelle issues de OpenStreetMap, des données concernant les quartiers, les espaces verts et les plages issues de l'Open Data de la Rochelle.

### 5.2. *Protocole d'expérimentation*

Le protocole de l'expérimentation est défini par les étapes suivantes : (1) collecter l'avis des experts sur la similarité ou la non-similarité de chaque paire de trajectoires de manière globale et selon chaque dimension ; (2) collecter les résultats issus de la mesure DA3T pour chaque paires de trajectoires de manière globale, selon chaque dimension ainsi que selon chaque niveau de granularité par dimension en ayant fixé les seuils (c.-à-d. valeur de la mesure au-delà de laquelle deux trajectoires sont considérées comme similaires) et coefficients ; (3) collecter les résultats issus des mesures DTW (Keogh, Ratanamahatana, 2005), de similarité de RI temporelle (Le Parc-Lacayrelle *et al.*, 2007) et MUITAS (May Petry *et al.*, 2019) correspondant respectivement aux mesures de référence spatiale, temporelle et thématique dans l'état de l'art ; (4) utiliser les métriques de précision (c.-à-d. nombre de paires évaluées similaires par la mesure et par les experts rapporté au nombre de paires évaluées similaires par la mesure mais pas forcément par les experts), de rappel (c.-à-d. nombre de paires évaluées similaires par la mesure et par les experts rapporté au nombre de paires évaluées similaires par les experts mais par forcément par la mesure) et de F1-mesure (c.-à-d. mesure combinant la précision et le rappel) pour calculer la pertinence de mesure DA3T par rapport à l'avis des experts puis la comparer avec les mesures de l'état de l'art ; (5) réitérer les étapes (2) et (4) avec des seuils et coefficients différents afin d'optimiser ces valeurs. Présentons maintenant les résultats de l'expérimentation.

### 5.3. *Résultats et discussion*

Dans un premier temps, nous décrivons notre travail relatif à l'optimisation des seuils et coefficients de pondération de la mesure DA3T. Dans un second temps, nous procédons à l'évaluation de notre mesure et commentons les résultats obtenus. Le tableau 2 présente les résultats issus de l'optimisation de la mesure DA3T faite consécutivement à la collecte des avis des experts. Il présente les scores micro, méso et macro en terme de rappel, précision et F1-mesure pour chaque dimension. Concernant la dimension spatiale, nous observons de légères disparités dans les résultats relatifs aux niveaux de granularité micro, méso et macro. Le niveau micro donne une F1-mesure de 0.889, légèrement supérieure aux deux autres. Le score spatial global prend en compte les trois niveaux de granularité de manière équivalente et nous constatons une amélioration des résultats par rapport à ceux des niveaux de granularité pris individuellement : F1-mesure à 0.914. Pour ce jeu de données, nous pouvons en conclure que les experts utilisent tous les grains dans leur observation d'une trajectoire touristique dans la ville.

TABLEAU 2. Résultats issus de l'exécution de la mesure DA3T

Score	$\alpha$	$\beta$	$\gamma$	Seuil	Précision	Rappel	F1-mesure
$S_{spt-mic}$	—	—	—	0.9	1	0.8	0.889
$S_{spt-mes}$	—	—	—	0.892	0.762	0.865	0.865
$S_{spt-mac}$	—	—	—	0.024	0.813	0.813	0.813
$S_{spt}$	0.33	0.33	0.34	0.616	1	0.842	0.914
$S_{tmp-mic}$	—	—	—	0.15	1	0.762	0.865
$S_{tmp-mes}$	—	—	—	0.22	0.813	0.929	0.867
$S_{tmp-mac}$	—	—	—	0.34	0.938	0.789	0.857
$S_{tmp}$	0.2	0.6	0.2	0.4	0.813	1	0.897
$S_{thm-mic}$	—	—	—	0.26	0.538	0.778	0.636
$S_{thm-mes}$	—	—	—	0.05	1	0.565	0.722
$S_{thm-mac}$	—	—	—	0.2	1	0.867	0.929
$S_{thm}$	0.2	0.1	0.7	0.275	0.923	0.857	0.889
$S_{glb}$	0.4	0.2	0.4	0.4	1	0.857	0.923

Concernant la dimension temporelle, nous observons que, parmi les trois niveaux de granularité micro, méso, macro, aucune ne se démarque des autres, toutes ont une F1-mesure autour de 0,86. Nous constatons une amélioration de cette métrique dans le score temporel global : F1-mesure à 0.897. Par conséquent, ici également, il est intéressant de considérer les trois niveaux de granularité pour calculer la similarité temporelle de deux trajectoires sémantiques.

Concernant la dimension thématique, nous observons de fortes disparités dans les résultats relatifs aux niveaux de granularité micro, méso et macro. Le niveau macro donne une F1-mesure de 0.929 nettement supérieure aux deux autres. Tout d'abord, nous en concluons que les experts privilégient les aspects dominants de chaque thématique (p. ex. météo globalement ensoleillée, visite centrée autour du quartier du port, activité de restauration prédominante, etc.). Notons également ici que le score thématique global obtenu n'améliore pas le score thématique macro : F1-mesure 0.889 et de 0.929 respectivement. Dans ce cas particulier, nous préconisons une pondération des coefficients micro, méso et macro à 0, 0 et 1 respectivement.

Enfin, les résultats obtenus avec la mesure DA3T globale sont supérieurs à ceux obtenus avec les mesures dimensionnelles considérées séparément, ce qui nous permet d'affirmer que les experts utilisent toutes les dimensions des trajectoires sémantiques pour les comparer. Nous constatons, cependant, que le coefficient de pondération attribué à la dimension temporelle et optimisant les résultats est un peu plus faible que ceux des dimensions spatiale et thématique, ce qui laisse à penser que les experts s'intéressent un peu moins à la dimension temporelle lorsqu'ils comparent deux trajectoires touristiques.

Passons maintenant à l'évaluation de notre mesure par rapport à des mesures de référence existantes dans les différentes dimensions et globalement. Les mesures de référence choisies sont : DTW présentée dans Keogh, Ratanamahatana (2005) pour la dimension spatiale, la mesure de similarité de RI temporelle présentée dans Le Parc-

TABLEAU 3. Comparaison de la mesure DA3T avec des mesures de référence grâce à la F1-mesure

Dimension	Mesure de référence			Mesure DA3T
	DTW	RI temp.	MUITAS	
Spatiale	0.889			0.914
Temporelle		0.867		0.897
Thématique			0.636	0.889
Combinées	0.857			0.923

Lacayrelle *et al.* (2007) pour la dimension temporelle et enfin la mesure MUITAS présentée dans May Petry *et al.* (2019) pour la dimension thématique. Concernant la mesure de référence combinant les trois dimensions, nous avons utilisé la moyenne des trois mesures précédemment citées. Le tableau 3 montre que la mesure DA3T donne des résultats s'approchant plus de l'avis des experts que des mesures de référence choisies, et ce, dans toutes les dimensions et globalement.

Pour conclure, nous reprenons les quatre questions annoncées en début de section. Premièrement, pour chaque dimension, nous obtenons une F1-mesure autour de 0.90, ce qui nous permet de dire que notre mesure donne des résultats relativement proches de l'avis des experts. Deuxièmement, nous avons optimiser les coefficients de pondération par grain et par dimension. Notons que, selon la dimension, les grains privilégiés sont différents. Troisièmement, les résultats de la mesure globale montrent qu'il est très pertinent de considérer les trois dimensions des trajectoires sémantiques pour les comparer. Enfin, quatrièmement, notre hypothèse de départ, qui, rappelons le, propose d'observer chaque dimension selon trois niveaux de granularité, est validée. En effet, les dimensions spatiale et temporelle montrent une amélioration du résultat global par rapport aux résultats granulaires pris individuellement. D'autres part, le système de pondération permet d'éviter d'éventuelles pertes de performance.

## 6. Conclusion

Cet article a permis de présenter une nouvelle fonction de calcul de similarité entre trajectoires sémantiques dans un cadre d'activités touristiques. Cette fonction prend en compte toutes les dimensions des trajectoires sémantiques (c.-à-d. spatiale, temporelle et thématique) et ce sur trois niveaux de granularité (c.-à-d. micro, méso et macro) offrant la possibilité d'analyses plus ou moins fines et précises selon les conditions et besoins exprimés. Chaque composante de notre mesure a un coefficient de pondération qui permet de paramétrer son importance dans le calcul final du score. Afin de valider l'efficacité de notre mesure, nous avons mis en place une expérimentation basée sur l'avis des experts qui s'est révélée concluante.

Dans de prochains travaux, nous souhaitons enrichir notre mesure en prenant en compte les aspects multi-dimensionnels des trajectoires (p. ex. la dimension spatio-temporelle pour tenir compte de la vitesse de déplacement). Nous souhaitons également mener une expérimentation de plus grande envergure avec plus d'experts et plus de couples

de trajectoires à évaluer. Actuellement générique, cette mesure peut également être améliorée afin de l'adapter à des contextes spécifiques (c.-à-d. déplacements humains touristiques ou professionnels, animaliers, etc.) et ainsi être pondérée de façon automatique selon ces contextes d'usage.

#### Remerciements

*Cet article a été écrit dans le cadre du projet DA3T, financées par la région Nouvelle-Aquitaine et la société Berger-Levrault.*

#### Bibliographie

- Aiello M. (2002, janvier). A spatial similarity measure based on games: Theory and practice. *Logic Journal of IGPL*, vol. 10.
- Allen J. F. (1983, novembre). Maintaining knowledge about temporal intervals. *Communications of the ACM*, vol. 26, n° 11, p. 832–843.
- Alt H. (2009, septembre). The Computational Geometry of Comparing Shapes. In *Efficient Algorithms: Essays Dedicated to Kurt Mehlhorn on the Occasion of His 60th Birthday*, p. 235–248. Berlin, Heidelberg, Springer-Verlag.
- Andrew J. (1805). Astronomical and nautical tables.
- Chen L., Ng R. (2004, janvier). On The Marriage of Lp-norms and Edit Distance. In, p. 792–803.
- Chen L., Özsu M. T., Oria V. (2005, janvier). Robust and fast similarity search for moving object trajectories. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, p. 491–502.
- Chen Y., Nascimento M. A., Ooi B. C., Tung A. K. H. (2007, avril). SpADe: On Shape-based Pattern Detection in Streaming Time Series. In *2007 IEEE 23rd International Conference on Data Engineering*, p. 786–795.
- Cleasby I. R., Wakefield E. D., Morrissey B. J., Bodey T. W., Votier S. C., Bearhop S. *et al.* (2019, novembre). Using time-series similarity measures to compare animal movement trajectories in ecology. *Behavioral Ecology and Sociobiology*, vol. 73, n° 11, p. 151.
- Egenhofer M. J. (1997, août). Query Processing in Spatial-Query-by-Sketch. *Journal of Visual Languages & Computing*, vol. 8, n° 4, p. 403–424.
- Faloutsos C., Ranganathan M., Manolopoulos Y. (1994, mai). Fast subsequence matching in time-series databases. *ACM SIGMOD Record*, vol. 23, n° 2, p. 419–429.
- Keogh E., Ratanamahatana C. A. (2005, mars). Exact indexing of dynamic time warping. *Knowledge and Information Systems*, vol. 7, n° 3, p. 358–386.
- Lee J.-G., Han J., Whang K.-Y. (2007, juin). Trajectory clustering: a partition-and-group framework. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, p. 593–604. New York, NY, USA, Association for Computing Machinery.

- Le Parc-Lacayrelle A., Gaio M., Sallaberry C. (2007). La composante temps dans l'information géographique textuelle. *Document Numérique*, vol. 10, n° 2, p. 129–148. (Publisher: Lavoisier)
- Lu E. H.-C., Tseng V. S. (2009, mai). Mining Cluster-Based Mobile Sequential Patterns in Location-Based Service Environments. In *2009 Tenth International Conference on Mobile Data Management: Systems, Services and Middleware*, p. 273–278.
- Magdy N., Sakr M., Abdelkader T., Elbahnasy K. (2015, décembre). Review on trajectory similarity measures.
- May Petry L., Ferrero C., Alvares L., Renso C., Bogorny V. (2019, juin). Towards semantic-aware multiple-aspect trajectory similarity measuring. *Transactions in GIS*, vol. 23.
- Mello R. D., Bogorny V., Alvares L. O., Santana L. H. Z., Ferrero C. A., Frozza A. A. *et al.* (2019, mai). MASTER: A multiple aspect view on trajectories. *Transactions in GIS*, p. tgis.12526.
- Moreau C., Devogele T., Etienne L. (2018, novembre). Extraction de motifs de trajectoires sémantiques similaires. In *Spatial Analysis and Geomatics*. Montpellier, France.
- Nakamura T., Taki K., Nomiya H., Seki K., Uehara K. (2013, novembre). A shape-based similarity measure for time series data with ensemble learning. *Pattern Analysis and Applications*, vol. 16, n° 4, p. 535–548.
- Parent C., Spaccapietra S., Renso C., Andrienko G. L., Andrienko N. V., Bogorny V. *et al.* (2013). Semantic trajectories modeling and analysis. *ACM Comput. Surv.*, vol. 45, n° 4, p. 42:1–42:32.
- Sallaberry C. (2013). *Geographical Information Retrieval in Textual Corpora*. Wiley-ISTE.
- Su H., Liu S., Zheng B., Zhou X., Zheng K. (2020, janvier). A survey of trajectory distance measures and performance evaluation. *The VLDB Journal*, vol. 29, n° 1, p. 3–32.
- Tao Y., Both A., Silveira R. I., Buchin K., Sijben S., Purves R. S. *et al.* (2021, juillet). A comparative analysis of trajectory similarity measures. *GIScience & Remote Sensing*, vol. 58, n° 5, p. 643–669.
- Varlamis I., Sardianos C., Bogorny V., Alvares L. O., Carvalho J. T., Renso C. *et al.* (2021, mars). A novel similarity measure for multiple aspect trajectory clustering. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*. New York, NY, USA, Association for Computing Machinery.
- Vlachos M., Kollios G., Gunopulos D. (2002). Discovering similar multidimensional trajectories. In *Proceedings 18th International Conference on Data Engineering*, p. 673–684. San Jose, CA, USA, IEEE Comput. Soc.
- Wang H., Su H., Zheng K., Sadiq S., Zhou X. (2013, janvier). An effectiveness study on trajectory similarity measures. In, p. 13–22.



---

# Innovation collaborative pour la mobilité des seniors

## Une démarche expérimentale de conception de services de covoiturage solidaire

**Christine Verdier**

Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France  
[christine.verdier@univ-grenoble-alpes.fr](mailto:christine.verdier@univ-grenoble-alpes.fr)

---

*RESUME. La modélisation des écosystèmes complexes est abordée dans la littérature par le biais de méthodes d'amélioration continue qui permettent à la fois de représenter les participants et leurs vues sur le système ainsi que les contraintes du domaine dans une approche centrée humain. Il n'en demeure pas moins que dès qu'un écosystème est dénué de procédures assez stables, se construit au fur et à mesure avec des services qui se créent à la volée, ces méthodes et les outils afférents comportent certaines limites. Nous proposons dans cet article une démarche expérimentale de conception de services de covoiturage solidaire pour les personnes âgées en milieu rural et périurbain (donc un milieu géographique contraint) basée sur une méthode d'amélioration continue étayée d'outils de scénarisation et de co-conception pour permettre une modélisation de l'écosystème le plus proche possible de la réalité et des contraintes du domaine.*

*ABSTRACT. The modeling of complex ecosystems is approached in the literature through continuous improvement methods that allow both to represent the participants and their views on the system and the constraints of the domain in a human-centered approach. The fact remains that as soon as an ecosystem is devoid of fairly stable procedures, is built gradually with services that are created on the fly, these methods and the related tools reach some limits. We propose in this article an experimental approach to the design of solidarity carpooling services for the elderly in rural and peri-urban areas (therefore a constrained geographical environment) based on a continuous improvement method supported by scripting and co-design tools to allow modeling of the ecosystem as close as possible to reality and the constraints of the field.*

*Mots-clés : écosystème, méthode d'amélioration continue, ingénierie des exigences, covoiturage solidaire*

*KEYWORDS: ecosystem, continual innovation method, requirement engineering, ridesharing*

---

## 1. Introduction

La mobilité des personnes âgées en zones rurales et périurbaines est un sujet multifactoriel dont les acteurs de l'écosystème concerné s'emparent de différentes manières en fonction de leur angle de vue. Les collectivités territoriales (communes, départements, régions), selon leurs niveaux de responsabilité respectifs, en font un argument majeur des politiques publiques pour permettre à ces personnes de rester citoyennes (en ayant accès à toutes les prérogatives afférentes) et actives (accession facile aux services marchands présents dans la commune). Les associations de seniors ainsi que les familles et les aidants cherchent des solutions pratiques quotidiennes pour ces personnes. Les personnes âgées elles-même ne sont pas toutes égales face à leur mobilité quotidienne. Certaines pratiquent un covoiturage spontané (entraide entre voisins, au sein de la famille, via des auxiliaires de vie), d'autres se tournent vers les centres communaux d'action sociale (CCAS) des communes, d'autres cherchent des solutions individuelles alternatives (marche, vélo), d'autres déménagent pour se rapprocher des centres-bourgs, d'autres encore se désocialisent progressivement pour devenir des « invisibles » de la mobilité. Force est de constater que les services de mobilité existants dans les communes bien que nombreux, fonctionnent assez mal. Les raisons sont multifactorielles et tiennent à la fois à la personne elle-même (refus de covoiturer avec quelqu'un d'autre, avec un conducteur qu'on ne connaît pas, horaires trop contraints, non connaissance des services disponibles, etc.), à la complexité d'utilisation du service, à l'inadéquation entre le service et le besoin et au modèle économique sous-jacent.

Les systèmes d'information apportent de nombreux outils et des méthodes pour modéliser les exigences, les processus métier, les contraintes et représentations des écosystèmes complexes et le covoiturage des personnes âgées en est un. Cependant, les contraintes du domaine sont telles qu'une démarche de conception centrée humain unique n'est pas adaptée.

Nous proposons dans cet article une démarche de co-conception de services de covoiturage basée sur une combinaison de fragments de méthodes centrées humain.

Après une introduction et un état de l'art, nous présentons la démarche expérimentale dont le fil conducteur est la méthode d'amélioration continue AdInnov (Cortes-Cornax *et al.*, 2015 ; 2016). Nous poursuivons par les résultats que nous avons obtenus puis quelques éléments sur l'expérimentation pour ouvrir ensuite la discussion sur les gains et limites de l'approche.

## 2. Etat de l'art

L'ingénierie des exigences est un courant scientifique largement utilisé pour modéliser les besoins des organisations. Les méthodes classiques d'ingénierie des exigences telles que Kaos (Van Lamsweerde, 2001) ou i\* (Castro *et al.*, 2002) proposent des langages graphiques de modélisation basés sur les buts. Dans (Wanderley *et al.*, 2014), les auteurs abordent le sujet sur le plan de la trop grande complexité à manipuler et comprendre des supports d'ingénierie des exigences par

des utilisateurs finaux. Ils proposent l'utilisation de modèles cognitifs qui seront alors transformés en modèles KAOS puis SYSML. D'autres recherches (Touzani *et al.*, 2016) montrent que l'apport de la géomatique peut améliorer et compléter l'ingénierie des exigences en localisant physiquement les objets physiques. Dans (Shambour *et al.*, 2022), les auteurs notent que la grande difficulté dans l'ingénierie des exigences réside dans leur élicitation et proposent dans les approches de conception logicielle des systèmes de recommandation semi-automatiques par l'intermédiaire de techniques intelligentes permettant d'améliorer cette phase. Dans un domaine aval, celui du développement logiciel, de nouvelles approches permettent d'associer des artefacts psychologiques pour mieux appréhender les connaissances à modéliser par les informaticiens (Graziotin *et al.*, 2022) ou intègrent des notions de résilience pour améliorer l'analyse des exigences dans les écosystèmes complexes avec une application aux systèmes d'information de santé (de Carvalho *et al.*, 2021).

Les méthodes d'amélioration continue sont largement diffusées dans la littérature (Arnheiter *et al.*, 2005 ; Sokovic *et al.*, 2010). L'une des plus populaires est le cycle Plan-Do-Check-Act (Deming, 2000). Assez peu sont bien adaptées à la modélisation des écosystèmes complexes. Nous utilisons dans notre recherche la méthode AdInnov (Cortes-Cornax *et al.*, 2015 ; 2016), (Front *et al.*, 2017) qui suit les principes de la roue de Deming et qui est une méthode d'amélioration continue participative pour l'analyse, le diagnostic et l'innovation des écosystèmes socio-techniques complexes.

La méthode ADInnov est issue de la généralisation de la méthode empirique suivie dans un précédent projet ANR Innoserv qui a été proposée et validée scientifiquement dans le domaine de l'ingénierie des méthodes. Elle est basée sur le cadre As-Is/As-If et a pour objectif d'imaginer des scénarios 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 nécessaires). Pour ce faire, elle repose comme toutes les méthodes issues du cadre As-Is/As-If, sur un cycle d'amélioration continue, par opposition aux approches projets ayant une équipe projet, un budget, une date de début et une date de fin, etc. Les évolutions sont organisées selon des roadmaps spécifiant quand et comment les déployer en fonction des contraintes juridiques, économiques, sociales ou techniques impactées par les évolutions proposées (Verdier *et al.*, 2018). Cette approche d'amélioration continue permet donc d'aborder les écosystèmes complexes par l'analyse de points de blocage dans les processus métiers actuels pour étudier comment la levée de ces points de blocage pourrait produire une amélioration du fonctionnement de l'organisation.

La méthode AdInnov est bien adaptée à notre contexte d'usage car la mobilité des personnes âgées en milieu rural et périurbain est par essence un écosystème complexe :

- Les acteurs et les rôles sont multiples : passagers, conducteurs, collectivités territoriales (avec chacune leurs prérogatives en matière de transport, de vieillissement, d'aménagement du territoire ou encore de financement), aidants, famille.
- Les services de mobilité proposés sont nombreux : transports en commun, transport solidaire, transport à la demande, mais très partiellement utilisés.

- La modélisation de l'écosystème par une gestion de projet traditionnel n'est pas adaptée car il est très difficile de trouver des « patterns » de comportements vis-à-vis de la mobilité quotidienne.

La méthode AdInnov comporte trois phases : analyse, diagnostic et innovation. Le résultat de la phase d'analyse produit notamment l'identification des acteurs et des enquêtes d'exploration, celui de la phase de diagnostic, l'identification des points de blocage et la modélisation du diagramme de buts (version simplifiée de Kaos). Le résultat de la phase d'innovation produit des innovations de services et des innovations organisationnelles.

La mobilité des personnes âgées est un sujet abordé très largement dans la littérature. Dans (Michel et Robié, 2013), les auteurs proposent une carte d'accessibilité pour les décideurs locaux en faveur de la mobilité des personnes âgées qui recense notamment les zones d'inclusion et d'exclusion. Dans (Mondou et Violier, 2010), les auteurs focalisent leur réflexion sur la corrélation entre le périurbain et le haut niveau de mobilité nécessaire. Ils montrent notamment que les politiques publiques de transport n'apportent pas de solutions satisfaisantes et que le regroupement de services en périphérie accroît encore le problème. Les personnes âgées cherchent des solutions alternatives basées sur la solidarité, voire le déménagement quand d'autres finissent par se désocialiser. Ce point est également ciblé par l'enquête menée par (Pochet et Corget, 2010) qui insistent sur les risques de « désadaptation » rencontrés par les retraités du périurbain lorsque l'accès à la voiture devient compliqué du fait de l'âge, le handicap ou le décès du conjoint.

De très nombreuses expérimentations sont mises en oeuvre sur le territoire avec plus ou moins de succès. Le site de France Mobilités<sup>1</sup> recense une quarantaine de projets ou d'expérimentations liées à la mobilité des seniors. Nous pouvons citer par exemple : Rezo Seniors (Rezo Pouce), SilverMobi, Vivocab, Rox Rox, Wimoov, Clem', IdvRoom, SoliMobi, DEFI Mobilité, CARL, ecov, Taxi à la carte, Andyamo, Mon Copilote, CAR - Conduire l'automobile du retraité -. Certaines expérimentations sont avancées et déjà bien ancrées dans les territoires et d'autres beaucoup plus en difficulté. Quelques éléments communs sont à mentionner : la grande difficulté à mobiliser dans le temps les conducteurs solidaires, le temps long d'acceptation du service par les personnes âgées ou encore la faible zone de chalandise (peu de conducteurs et personnes âgées concernés) malgré un besoin très fort et individualisé.

Notre positionnement se place dans le cadre de la méthode AdInnov et une démarche expérimentale de co-conception basée sur un assemblage de méthodes pour réussir à contourner les problèmes majeurs de cet écosystème : difficulté à formaliser des services de covoiturage qui fonctionnent « en routine », difficulté à intégrer des contraintes nombreuses et fluctuantes de l'écosystème. Notamment, l'approche empirique abordée dans ce projet, à partir du cadre AsIs/AsIf a mis en lumière la nécessité de mieux formaliser les innovations de services (cf. par. Discussion).

---

<sup>1</sup> [www.francemoblites.fr](http://www.francemoblites.fr)

### 3. Démarche expérimentale

#### 3.1. Description du projet Mobipa

Le projet Mobipa – Mobilité Inclusive pour les Personnes âgées – financé par la région Auvergne Rhône-Alpes<sup>2</sup> a pour but de concevoir et mettre en œuvre des services de mobilité basés sur le covoiturage solidaire. Dès qu'un senior est en situation de fragilité physique (n'ose plus conduire, a des difficultés à marcher, etc.), sociale (est isolé) ou encore numérique (a des difficultés à utiliser l'outil numérique, à naviguer sur internet ou encore n'a pas accès à un ordinateur), il se trouve de fait dans la difficulté d'avoir accès à des services en ligne, à se déplacer pour faire des courses ou même se rendre à un rendez-vous médical. Le but du projet Mobipa est donc de permettre à ces personnes de pouvoir se déplacer facilement par des solutions de covoiturage afin de continuer à être des citoyens à part entière. Le cadre de MOBIPA s'inscrit dans le covoiturage solidaire et intergénérationnel pour venir en aide au public fragile et pour faire se rencontrer des publics souvent éloignés par l'âge, l'activité ou les relations interpersonnelles. Il se base sur une approche low tech<sup>3</sup> de co-conception de services de covoiturage avec les différentes parties prenantes.

#### 3.2. Démarche de co-conception du service Mobipa

##### 3.2.1. Cadre de référence

Nous avons développé notre démarche dans le cadre de la méthode AdInnov qui est une méthode pour analyser, diagnostiquer et proposer des innovations dans les écosystèmes complexes. La Figure 1 représente la méthode AdInnov avec le formalisme MAP (Dimassi *et al.*, 2008). Ce formalisme représente le niveau intentionnel de l'écosystème et permet de visualiser le plus haut niveau du SI actuel (As-Is) et futur (As-If) de mobilité. Les nœuds représentent les intentions (<caractériser>, <imaginer>) et les arcs les stratégies.

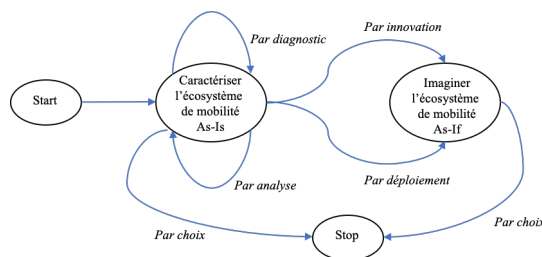


Figure 1: MAP de la méthode AdInnov appliquée à Mobipa

<sup>2</sup> Projet Pack Ambition Recherche 2018

<sup>3</sup> Nous entendons par « low tech » les outils numériques disponibles au domicile et avec lesquels la personne âgée est à l'aise. Il peut s'agir uniquement d'un téléphone fixe.

Les stratégies ont été définies sur la base de plusieurs cadres méthodologiques : cadre global emprunté au design thinking (Brown, 2009) qui a mobilisé en renfort des outils méthodologiques emprunté au Scenario Based Design (Rosson et Carroll, 2002) et (Forest *et al.*, 2009), à l'innovation par les usages (Mallein et Toussaint, 1994) et (Forest *et al.*, 2013), et à l'innovation sociale (Artis *et al.*, 2020).

Le Design Thinking peut se définir comme “un mode d'application des outils de conception utilisés par les designers pour résoudre une problématique d'innovation, par une approche multidisciplinaire centrée sur l'humain”<sup>4</sup>. Le processus Design Thinking stabilisé et proposé aujourd'hui dans la littérature (Caron-Fasan et Zerbib, 2017), se décompose en cinq phases : Empathize, Define, Ideate, Prototype, Test.

L'approche Scenario-by-Design (Rosson et Carroll, 2002) a pour paradigme d'intégrer des scénarios d'interaction utilisateur pour intégrer des pratiques d'utilisabilité dans le développement des systèmes interactifs. L'approche permet donc de rendre réaliste la solution future lors de la phase de conception. Nous l'avons adapté sur le fond : pour représenter le service de covoiturage idéal (mais facilement opérationnel) ; et sur la forme en utilisant des techniques de dessin. Le tableau suivant (Tableau 1) présente les fragments des différentes méthodes utilisées dans chaque étape de la démarche expérimentale du projet et réparties dans les phases d'analyse, de diagnostic et d'innovation de la méthode AdInnov.

Les étapes sont les suivantes : Une première phase d'exploration pour mieux comprendre les enjeux de mobilité des personnes âgées en milieu rural, identifier les acteurs concernés et caractériser les premières hypothèses de solutions ; une deuxième phase de conception participative avec les acteurs des territoires pour le développement d'un service innovant qui fait sens pour eux ; enfin une phase d'évaluation du service auprès des différents profils d'utilisateurs afin de valider la pertinence du service pour les parties prenantes.

Tableau 1: Méthodes utilisées dans les phases du projet Mobipa

<b>AdInnov : PHASE D'ANALYSE (As-Is)</b>			
<b>Activités</b>	<b>Etapes Mobipa</b>	<b>Fragment méthode</b>	<b>Moyens utilisés</b>
Ciblage acteurs intermédiaires <sup>5</sup>	Exploration	Design Thinking (E/D)	Entretiens communes Etude littérature
Caractérisation offre de services existants	Exploration	Design Thinking (E/D)	Etude littérature Etude solutions existantes de covoiturage

<sup>4</sup> Source:<http://www.frenchweb.fr/le-design-thinking-un-nouvel-avantage-competitif/122936>, 2017-01-13, repéré le 18 novembre 2021

<sup>5</sup> Les acteurs intermédiaires sont en relation étroite avec les personnes âgées : directeur/trice d'Ehpad, médecin généraliste, responsable de CCAS, maire, services à la personne, etc. Ils participent à l'écosystème.

			Entretiens semi-directifs
Définition cible et critères d'inclusion	Exploration	Design Thinking (I)	Brainstorming Entretiens semi-directifs d'acteurs-clés
<b>AdInnov : PHASE DE DIAGNOSTIC (As-Is)</b>			
Définition points de blocage et diagramme de buts	Exploration	Design Thinking (I)	Entretiens semi-directifs Focus group Etude des solutions existantes Modélisation conceptuelle
Contraintes et opportunités	Exploration	Design Thinking (I)	Entretiens des acteurs-clés Enquêtes personnes âgées
<b>AdInnov : PHASE D'INNOVATION (As-If)</b>			
Co-conception de services	Conception participative	Scenario-by-Design et Innovation par les usages	3 ateliers de co-création avec des acteurs-clés et des personnes âgées. Représentation physique du service de covoiturage sur un fond de carte de territoire Définition de 3 scénarios possibles réalistes.
Test des services	Evaluation	Innovation sociale et innovation par les usages	Enquêtes d'usage et focus-group

La phase d'innovation est détaillée dans le paragraphe 4.

### 3.2.2. Phases d'étude de l'écosystème

#### 3.2.2.1. Analyse de l'écosystème de mobilité

La phase d'analyse est présentée dans la Figure 2.

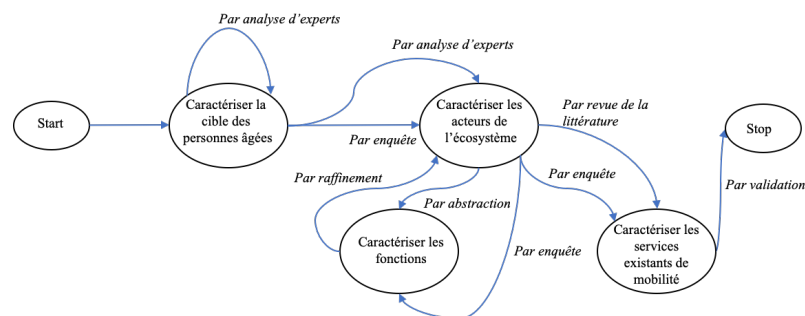


Figure 2 : Analyse de l'écosystème de mobilité

Cette phase a permis de mieux caractériser le public cible principal auquel s'adresse le service à concevoir (Verdier *et al.*, 2019). Cela a abouti à la définition de critères d'inclusion : personnes n'ayant plus d'activité professionnelle, en autonomie financière, en capacité de marcher, disposant d'un véhicule qu'elles utilisent peu ou qu'elles ne conduisent pas ou plus, ne disposant pas de véhicule, ayant un accès à un téléphone ou un smartphone ou une connexion internet, souhaitant avoir des activités sociales à l'extérieur du domicile et n'ayant pas d'aidant proche pour la mobilité quotidienne. Ont été exclus de cette cible les personnes âgées ayant un handicap nécessitant un véhicule médicalisé et des accompagnants professionnels.

Un dictionnaire des termes, une définition des acteurs et des rôles ont été définis à l'issue de cette phase mais ne sont pas présentés ici.

### 3.2.2.2. Diagnostic de l'écosystème de mobilité

La phase de diagnostic est présentée dans la figure suivante (Figure 3).

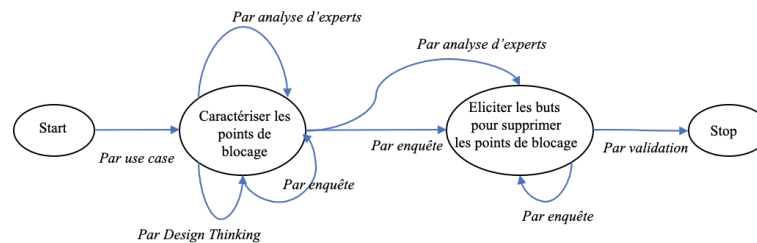


Figure 3: Diagnostic de l'écosystème de mobilité

Les points de blocage ont mis en exergue la nécessité de construire des services de confiance (que ce soit au niveau des prescripteurs : tiers de confiance institutionnels – médecin, CCAS, etc.-, au niveau des conducteurs ou au niveau du service lui-même), des services utilisables à la volée, qui permettent l'organisation de bout en bout du trajet (domicile-domicile) et dont les modalités financières sont simples et acceptables. Ils sont représentés dans le diagramme de buts suivant (Figure 4).



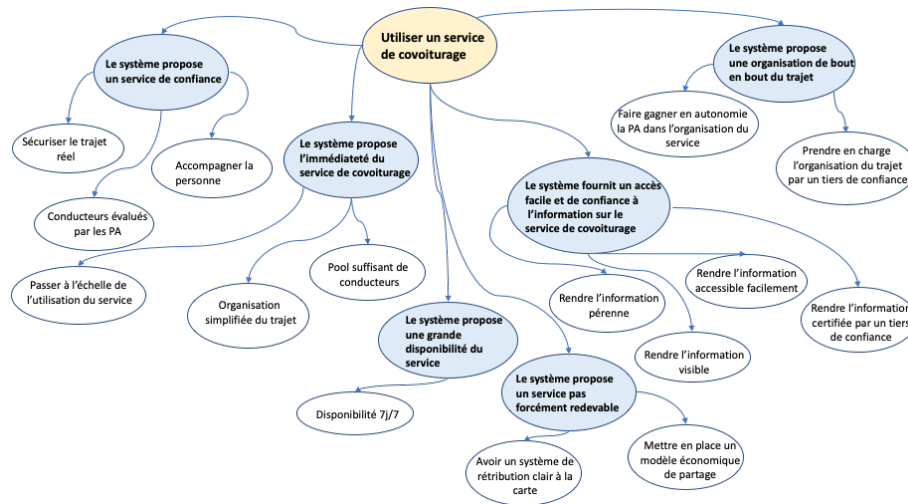


Figure 4: Diagramme de buts

### 3.2.2.3. Innovations de services de l'écosystème de mobilité

Les innovations de services sont présentés dans la figure suivante (Figure 5).

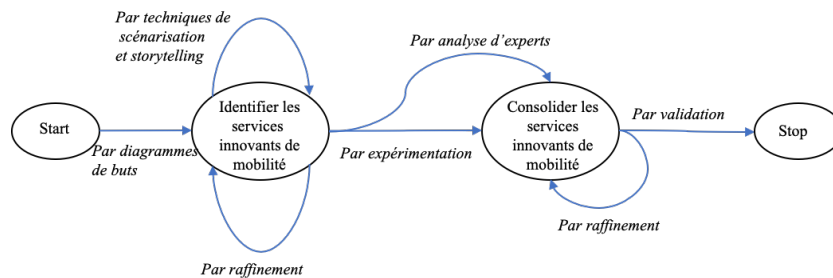


Figure 5: Innovations de l'éco-système

La session de co-création des services de mobilité s'est déroulée en trois temps d'atelier. Les participants étaient les suivants : trois membres de CCAS, un médecin généraliste, une orthophoniste, trois personnes âgées, cinq membres du consortium. L'atelier a été organisé et mené par trois membres de la Cité du Design. Les participants ont été répartis sur trois tables pour travailler sur trois communes situées dans les environs de Grenoble (Vif, Vizille et Lalley).

Temps 1 : L'objectif était de constituer une cartographie des scénarios de mobilité des personnes âgées. Ces cartes ont servi à identifier les différents acteurs impliqués dans ces mobilités et les problèmes rencontrés par les personnes âgées. Les trois

groupes représentant les trois communes ont travaillé sur trois types de déplacements : social (événement familial, sortie culturelle, etc.), médical (consultation de suivi, rendez-vous hospitalier, etc.), pratique (courses au supermarché, etc.).

Temps 2 : L'objectif était d'imaginer des services de mobilité collaboratifs nouveaux qui répondent aux problèmes identifiés lors du premier temps d'atelier. Les participants ont été invités à imaginer des solutions pour améliorer les services de mobilité des personnes âgées en milieu péri-urbain et rural avec la contrainte d'un service collaboratif et partagé.

Restitution : Les différents scénarios produits au cours de l'atelier ont été présentés par chacun des groupes et placés sur une courbe de complexité (Figure 8) permettant de juger de la faisabilité et de la probabilité de mise en œuvre du service.

Nous présentons ci-dessous les résultats pour la ville de Vif.

Les participants ont utilisé le fond de carte (Figure 6)<sup>6</sup> de la ville pour imaginer un covoiturage relatif au premier déplacement (social).



Figure 6: Fond de carte de Vif

Le scénario imaginé par ce groupe de participants a été intitulé : *le maillon manquant* : un service qui fait se rencontrer l'offre et la demande. Ce service permet aux personnes âgées de se déplacer à plusieurs en centralisant leurs besoins et leurs emplois du temps et en leur proposant de mutualiser leurs déplacements.

Déroulement — La personne âgée contacte la plateforme (cf par. 4.4.). Pour transmettre sa demande (type, date, horaire du déplacement) et ses besoins (aller-

<sup>6</sup> Les figures (Figure 6, Figure 7, Figure 8, Figure 9) sont l'œuvre de la Cité du Design (<https://citedudesign.com/fr/>)

retour, mobilité réduite, accompagnement, etc.). Elle précise le canal par lequel elle souhaite être recontactée (sms, appel, mail, etc.). L'opérateur du service rappelle la personne âgée pour lui indiquer les options qu'elle a trouvées pour son déplacement et convenir d'un rendez-vous. Le jour même, le déplacement a lieu avec d'autres personnes âgées qui avaient une demande similaire. Quelque soit le moyen de transport suggéré, l'accent est mis sur l'accompagnement et le collectif. Une possibilité a été émise de développer un réseau d'entraide et de partage plus large que le simple partage de déplacements. Ce scénario est représenté dans la figure (Figure 7)

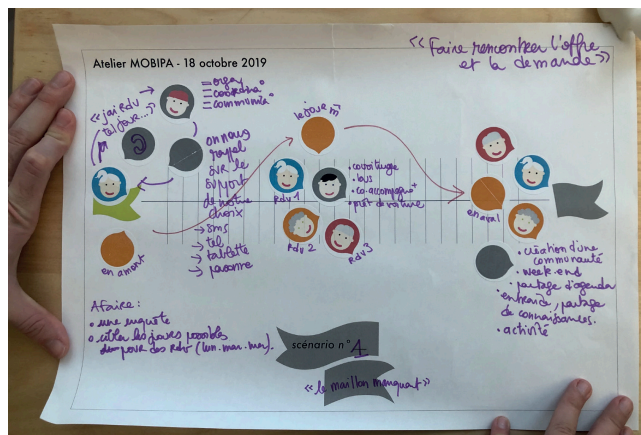


Figure 7 : Scénario Vif

Deux autres scénarios ont été proposés : un scénario « Mobil'Ainé » basé sur l'individualité et le porte-à-porte (Vizille) médié par une solution technologique (smartphone ou plateforme web) et un troisième scénario (Lalley) dont l'accent a été mis sur la mise en contact par une personne physique et sans médiation numérique.

D'autres solutions de mobilité ont été également étudiées durant ces ateliers : la conduite accompagnée de jeunes apprenants par des seniors, l'auto-école intergénérationnelle et solidaire, le covoiturage inversé (utilisation par des actifs du véhicule inutilisé d'un senior) et l'utilisation de flux pendulaires des actifs.

Tous ces scénarios et propositions ont été replacés sur une courbe de faisabilité (Figure 8).

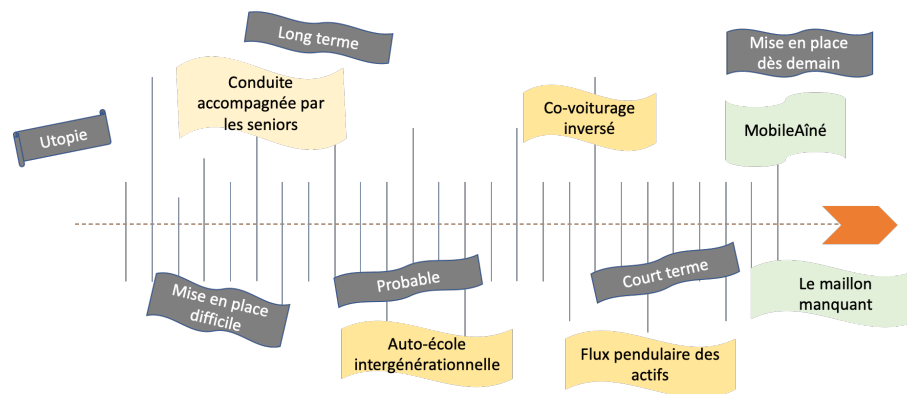


Figure 8 : Courbe de complexité et faisabilité

#### 4. Résultats obtenus

##### 4.1. Innovations de services

Les services dans AdInnov ont été représentés dans un métamodèle (Cortes-Cornax *et al.*, 2015 ; 2016), (Front *et al.*, 2017) avec une instanciation graphique qui contenait : le nom du service, le rôle et/ou l'acteur concerné, ainsi que les préoccupations<sup>7</sup> du service (financière, médicale, etc.).

Le service Mobipa est représenté dans le tableau suivant.

Tableau 2: Définition du service Mobipa

Modalités du service	Service simple : Déplacement Service complexe : Déplacement + activité
Activités	But du déplacement. Peut être simple ou complexe.
Type de service	Type générique à instancier sur chaque territoire
Moyen d'accès au service	Numérique, médié ou direct

##### 4.1.1. Déplacement et modalités du service

Le rôle du conducteur est porté par toute personne inscrite dans la plateforme. Les conducteurs envisagés dans le projet sont : des personnes récemment retraitées, des personnes âgées (disposant d'une voiture et conduisant), des jeunes actifs réalisant

<sup>7</sup> Les préoccupations qui représentent des vues sur le service (médical, social, financier, etc.) n'ont pas été encore prises en compte dans Mobipa.

des trajets pendulaires, des personnes au chômage (pouvant utiliser le véhicule de la personne âgée passagère), des jeunes apprentis conducteurs (lors de sessions de conduite avec une auto-école). Le rôle du passager est porté par la personne âgée. L'aller et le retour peuvent être réalisés par des conducteurs différents. Il s'agit d'un aller-retour domicile-domicile et les trajets de longue durée sont exclus au profit des trajets de la vie quotidienne.

Le service Mobipa a deux modalités : un service simple qui représente la tâche de déplacement et une activité simple ; ex : véhiculer une personne âgée de son domicile au marché, voyage aller-retour ; un service complexe composé d'une tâche de déplacement et d'une activité complexe ; ex : véhiculer une personne âgée de son domicile au supermarché, l'accompagner dans le supermarché et l'aider à porter ses courses et voyage retour.

#### *4.1.2. Activité*

L'activité représente le but à atteindre. Elle peut être composée ou non. Si elle est composée, les sous-activités peuvent être interdépendantes et doivent être cohérentes les unes par rapport aux autres (absence de conflit entre les sous-activités). La nature de l'activité n'a pas d'importance.

#### *4.1.3. Type de service*

Le service a été pensé de manière suffisamment générique pour permettre aux territoires souhaitant expérimenter le dispositif de l'instancier en fonction des critères locaux : géographie du territoire, profils des habitants âgés ou encore solutions de transports en commun disponibles.

#### *4.1.4. Moyens d'accès au service*

Les spécifications fonctionnelles d'une plateforme web de covoiturage ont été réalisées. Cependant, l'utilisation peut être directe via un site web ou une application mobile mais aussi médiée (par un personnel de la mairie, un voisin, la famille, etc.).

## **4.2. Fonctionnement global**

### *4.2.1. Scénarios*

Deux scénarios ont été construits suite aux ateliers de co-conception : Denis à Vif qui se rend à une visite médicale et Emilia à Vizille qui souhaite se rendre chez le coiffeur. Le scénario Emilia est présenté ci-après (Tableau 3).

Tableau 3: Scénario Commune Vizille

<b>Scenario 2 : Rendez-vous d'Emilia chez le coiffeur</b>		
<b>Description :</b> Emilia habite la commune de Vizille. Elle souhaite se rendre chez le coiffeur à 10h30, dans la commune de Champs-sur-Drac. Le trajet d'une dizaine de minutes en voiture prend 30 minutes en bus et 1h20 à pied.		
<b>Avant</b>	Etape 0	Emilia est une habituée de la plateforme Mobipa qu'elle utilise régulièrement depuis un an (1 à 2 fois par semaine). Elle y accède via l'application installée sur sa tablette lors d'un atelier organisé par le CCAS de Vizille. Depuis quelques temps, elle a commencé à proposer, via l'application, du soutien scolaire en mathématiques, matière qu'elle a enseignée pendant plus de 30 ans.
	Etape 1	A 8h30, elle se connecte sur l'application Mobipa. Elle indique l'adresse et le lieu de rendez-vous puis confirme son choix. Par défaut, l'adresse de départ est forcément l'adresse de son domicile qu'elle a renseignée la première fois. Emilia voudrait aller chez le coiffeur puis au marché et ne sait pas à quelle heure elle souhaite rentrer, elle ne choisit donc qu'un trajet aller. La plateforme cherche des possibilités de trajet dans sa base de données. L'auto-école Centr'Auto Formation de Vizille a justement des heures de conduite programmées entre 9h et 11h ce jour-là.
	Etape 2	A 8h45, Emilia reçoit une notification sur son iPad lui indiquant que la voiture de l'auto-école Centr'Auto Formation conduite par Zoé, accompagnée de Rachid, sera devant son domicile à 10h. L'auto-école reçoit également une notification lui indiquant que le formateur doit prévoir son trajet en passant par le domicile d'Emilia vers le salon de coiffure à 10h. Les numéros de téléphone d'Emilia et de l'auto-école sont échangés.
<b>Pendant</b>	Etape 3	A 10h, le véhicule de l'auto-école attend Emilia devant son domicile (10 minutes d'attente maximum). En cas de besoin, le moniteur appelle Emilia.
	Etape 4	Le trajet entre le domicile et le coiffeur a lieu.
	Etape 5	Emilia arrive devant le coiffeur à 10h20 et prend le temps de boire un thé en attendant son tour. Rachid, le moniteur de l'auto-école indique sur son application Mobipa que le trajet a eu lieu.
	Etape 6	Après son rendez-vous chez le coiffeur, Emilia se rend au marché. Son heure de retour étant imprévisible, elle n'a pas réservé son trajet en amont. Lorsqu'elle a terminé ses courses, elle se rend à la borne interactive Mobipa, située sur la place du marché (ou dans la mairie), elle peut y réserver un trajet retour vers son domicile. La borne lui indique qu'une conductrice, Camille, est disponible et sera là dans 10 minutes.
	Etape 7	Camille se gare devant la borne et se présente devant Emilia. Elle la conduit jusqu'à son domicile à Vizille.
<b>Après</b>	Etape 8	Le soir en rentrant chez elle, Emilia voit sur sa tablette une notification de l'application Mobipa lui demandant d'indiquer si les trajets ont eu lieu et lui proposant de donner son avis sur ces trajets.

#### 4.2.2. *Processus global du service*

Le service de mobilité se déroule en 8 étapes :

Etape 0 — Découverte de la plateforme Mobipa de covoiturage — A ce niveau, les personnes âgées et/ou les acteurs intermédiaires (tiers de confiance : CCAS, mairie, aidants) s'approprient l'utilisation de la plateforme numérique.

Etape 1 — Accès à la plateforme — La personne âgée ou le tiers de confiance recherche des conducteurs susceptibles de correspondre au besoin de trajet. La plateforme intègre un modèle multi-agent susceptible de faire du matching automatique entre les contraintes des personnes âgées et des conducteurs et de produire une liste de candidats potentiels (cf Figure 10).

Etape 2 — Mise en relation — La sélection du conducteur se fait de manière semi-automatique (par la plateforme et la personne âgée et/ou le tiers de confiance).

Etape 3 — Prise en charge — Le conducteur est prévenu et se rend chez la personne âgée à la date et l'heure convenues.

Etape 4 — Trajet aller — La personne âgée est véhiculée par le conducteur jusqu'au lieu de rendez-vous.

Etape 5 — Arrivée — Le conducteur accompagne à pied la personne âgée jusqu'à la porte du lieu de rendez-vous.

Etape 6 — Trajet retour — Le trajet retour est effectué par le même conducteur ou par un conducteur différent que la personne sollicite via son smartphone, le téléphone de son lieu de son rendez-vous ou encore des bornes interactives placées à proximité de son lieu de rendez-vous.

Etape 7 — Retour au domicile — Le conducteur accompagne à pied la personne âgée jusqu'à la porte de son domicile.

Etape 8 — Evaluation du service — La personne âgée fait une évaluation via l'application ou son téléphone du service. L'évaluation du service prend alors la forme de collecte de points donnant accès à des biens ou services dans les entreprises de commerce locales.

La représentation de ce processus global de covoiturage est mentionnée dans la figure suivante (Figure 9).



Figure 9: Processus global du service de covoiturage Mobipa

Un film de présentation du service de covoiturage Mobipa est accessible sur le lien suivant :

<https://drive.google.com/file/d/1HLJc5AFuINVY4Jtro2gz-y9HK3njuWqZ/view>.

#### 4.3. Expérimentations

Des expérimentations ont jalonné le projet.

En amont du projet, plusieurs focus group et entretiens collectifs et individuels ont été organisés sur les trois terrains de référence. Ils ont contribué à l'aboutissement de l'analyse et du diagnostic. Des enquêtes ont également été menées par le partenaire WeTechCare (filiale d'Emmaüs Connect) sur l'inclusion numérique des seniors qui a notamment montré la nécessité de construire une solution technologique qui soit directement utilisable par les personnes âgées si leur niveau d'inclusion numérique était fort (utilisation courante du smartphone par exemple) mais aussi médié dans le cas contraire. Des focus group ont également été menés avec les personnes âgées relativement à leur niveau et exigence de mobilité. Cette partie a été partiellement freinée par la pandémie de Covid 19. Des entretiens semi-directifs ont été menés avec des conducteurs.

En aval du projet, des enquêtes ont été menées pour évaluer le concept Mobipa, c'est-à-dire les ressentis de la solution conceptualisée. Les enquêtes ont montré que le service n'apparaissait pas comme différenciant par rapport à des solutions classiques



de covoiturage. Un réajustement a alors été fait dans la solution conceptuelle selon trois axes :

- La distinction entre covoiturage et covoiturage solidaire. Deux profils de conducteurs ont été proposés : des conducteurs habilités (habilitation délivrée par la mairie) et des conducteurs non habilités (avec médiation par un tiers de confiance).
- L'amélioration de la confiance. Trois propositions ont été faites : l'organisation du trajet de bout en bout y compris l'organisation des étapes intermédiaires en cas de trajet indirect, la gestion des « derniers mètres » en assurant le porte à porte et l'identification pour les conducteurs d'une place de parking libre la plus proche (une petite expérimentation en utilisant un réseau LoRa a été menée).
- Le renforcement du côté pratique. Trois propositions ont été faites :
  - Un service de covoiturage à la carte : modalités de réservation directes ou médiées, accompagnement possible, modalités diverses (service simple ou complexe) ;
  - Un service de confiance : validé et organisé par un tiers de confiance en minimisant le temps d'attente, un matching entre conducteurs et passagers ;
  - Un service organisé de bout en bout : existence d'une solution dégradée en cas de défaillance du conducteur, évaluation du service.

#### **4.4. Spécification de la plateforme d'intermédiation Mobipa**

Une plateforme de covoiturage met en relation des conducteurs et des passagers qui font le même trajet. Certaines particularités liées au projet Mobipa complexifient l'utilisation de la plateforme : les conducteurs n'ont pas de raison de faire le trajet que veulent faire les passagers, un accompagnement est parfois nécessaire, le service est multicanal, la livraison du service peut être médiée (par des intermédiaires de confiance), la composition d'un couple (conducteur, passager) est multicritère. Une ébauche de système multi-agent<sup>8</sup> (Lohja *et al.*, 2020a ; 2020b) a été réalisée afin de créer un matching automatique entre les conducteurs et les passagers. Dans la figure (Figure 10), la négociation est formalisée à l'aide d'un *Contract Net Protocol*. Dans la partie de droite (a), la proposition du trajet est formalisée. Elle contient la spécification du trajet, les contraintes d'éligibilité et les spécifications du contrat de covoiturage. La partie gauche (b) correspond à la spécification du contrat par un conducteur et la partie (c) correspond à la contractualisation.

---

<sup>8</sup> Cette partie multiagent n'a pas pu être finalisée.

<p><b>To:</b> *</p> <p><b>From:</b> Manager (for Operator)</p> <p><b>Type:</b> RIDE ANNOUNCEMENT</p> <p><b>Contract:</b> 29-02-2020</p> <p><b>Task Abstraction:</b> share ride from V to C round trip accompaniment</p> <p><b>Eligibility Specification:</b> must have proposed a ride must be leaving from or through V must be going to or through C must arrive in C around 9:00</p> <p><b>Bid Specification:</b> origin of the ride destination of the ride time of arrival at destination willingness to accompany gender</p> <p><b>Expiration time:</b> 28 February 2020, 23:59</p> <p><i>(a) Ride announcement.</i></p>	<p><b>To:</b> Manager (for Operator)</p> <p><b>From:</b> Contractor (for Pauline)</p> <p><b>Type:</b> BID</p> <p><b>Contract:</b> 29-02-2020</p> <p><b>Node Abstraction:</b> origin is V destination is C arrival at 9:00 round trip I am female I can accompany</p> <p><i>(b) Ride bid.</i></p> <p><b>To:</b> Contractor (for Pauline)</p> <p><b>From:</b> Manager (for Operator)</p> <p><b>Type:</b> AWARD</p> <p><b>Contract:</b> 29-02-2020</p> <p><b>Task Specification:</b> origin is V destination is C arrival at 9:00 round trip accompaniment</p> <p><i>(c) Ride contract award.</i></p>
--	--

Figure 10: Contractualisation d'un service de covoiturage

## 5. Discussion et conclusion

Ce projet de recherche-action a mis en lumière essentiellement deux éléments.

La co-construction et l'assemblage de différents fragments de méthodes pour modéliser cet écosystème particulièrement complexe a montré que le service de covoiturage réalisé correspondait aux besoins et représentations des personnes âgées et des acteurs intermédiaires. Pour aller plus loin, un atelier de restitution qui se déroulera en avril 2022 étudiera, sous forme de jeu sérieux entre les membres du consortium et les acteurs intermédiaires des territoires, les possibilités d'instancier tout ou partie des services proposés dans le projet (à partir des deux cas d'utilisation).

Concernant la méthode AdInnov, la modélisation du service a montré quelques limites. En effet, le projet a mis en exergue la notion de modèle d'affaire lié au covoiturage solidaire. Se pose donc la question de la valeur du service en fonction du rôle de l'acteur dans le SI. Nous réfléchissons à enrichir le modèle de classe et la MAP en ajoutant des éléments liés à la valeur du service. Cette valeur peut être financière, morale ou qualitative. Plus largement, la méthode mérite d'être complétée par l'intégration d'autres modèles. Outre les modèles d'affaire, le projet a également mis en exergue des différences liées aux lieux et typologie des espaces géographiques de vie que nous réfléchissons également à intégrer dans la méthode.

### Remerciements

*L'autrice remercie la région Auvergne Rhône-alpes pour le financement du projet Mobipa (Pack Ambition Recherche) ainsi que les membres du consortium pour leur apport scientifique dans le projet.*

## Bibliographie

- Arnheiter E. D., Maleyeff, J. The integration of lean management and Six Sigma. *The TQM Magazine* (17:1), pp. 5-18, 2015. (doi: <https://doi.org/10.1108/09544780510573020>).
- Artis A., Ribeiro L. Enquête d'usages, de coopération interorganisationnelle et modèles économiques du covoiturage. Livrable 3. Avril 2021.
- Brown T. *Change by design: how design thinking transforms organizations and inspires innovation*. Harper Collins Publishers. 272p., 2009.
- Caron-Fasan ML., Zerbib O. Réapprendre à s'étonner et à innover avec le design thinking. in *The Conversation*, 3 octobre 2017.
- Castro J., Kolp M., Mylopoulos J. Towards requirements-driven information systems engineering: the Tropos project. *Inf. Syst.*, vol.27, n°6, pp: 365-389, 2002.
- Cortes-Cornax M, Rieu D., Verdier C., Front A., Forest F., Mercier A., Benoit A.M., Faravelon A. A Method to Analyze, Diagnose and Propose Innovations for Complex Ecosystems: the InnoServ Project. *AHA 2015 – ER-Workshop on Conceptual Modeling for Ambient Assistance and Health Ageing*. Stockholms.
- Cortes-Cornax M., Front A., Rieu D., Verdier C., Forest F. ADInnov: A Method to Instil Innovation in Socio-technical Ecosystems. *Caise 2016* :133-148, Ljubljana, Slovenia, June 2016.
- de Carvalho E.A., Gomes J.O., Jatobá A. et al. Employing resilience engineering in eliciting software requirements for complex systems: experiments with the functional resonance analysis method (FRAM). *Cogn Tech Work* 23, 65–83 (2021). <https://doi.org/10.1007/s10111-019-00620-0>.
- Deming W. E. 2000. *The New Economics for Industry, Government, Education*. 2000. MIT Press.
- Dimassi J., Rolland C., Kraeim N. Le formalisme MAP. L'ingénierie des méthodes au service des nouvelles tendances de développement des applications informatiques. *Centre de Publication Universitaire de Tunisie*, pp.380, 2008. ([hal-00706104](https://hal.archives-ouvertes.fr/hal-00706104))
- Forest F., Chanal V., Lavoisy O. Integrated Scenario-based Design Methodology for Collaborative Technology Innovation. *The Future of Innovation (ISPIM - International Journal of Innovation Management)*, Jun 2009, Vienne, Austria. pp.99. halshs-00417935
- Forest F., Arhippainen L., Mallein P. Paradoxical User Acceptance of Ambient Intelligent Systems – Sociology of User Experience Approach. *Mindtrek 2013*. Tampere. Finland.
- Front A., Rieu D., Cela O., Movahedian F. Les méthodes d'évolution continue au sein des organisations : le cadre As-Is/As-If. *Inforsid 2017*, p.311-326.
- Graziotin D., Lenberg P., Feldt R., Wagner S. Psychometrics in behavioral software engineering: a methodological introduction with guidelines. *ACM Transactions on Software Engineering and Methodology*. Volume 31, Issue 1, January 2022. Article No 7 p 1-36. <https://doi.org/10.1145/3469888>
- Lohja I., Demazeau Y., Verdier C. A multi-agent system approach to ridesharing for older people: state-of-the-art work and preliminary design. *RJCLIA 2020*, Angers, 29 juin-3 juillet 2020.

- Lohja I., Verdier C., Front A. Towards a utilized ridesharing service for older people: a new approach. Short paper. *ICIS 2020*, 9 pages. Virtual conference.
- Mallein P., Toussaint Y. Technologies de l'information et de la communication : sociologie pour la Conception Assistée par l'Usage. *Communications&Stratégies n°5*. Cahiers de l'IDATE, 1994.
- Michel B., Riobé A.L. La carte d'accessibilité, un outil au service des décideurs locaux en faveur de la mobilité des personnes âgées. *ESO Travaux et Documents, volume 13*. 2013, p. 59-66.
- Mondou V., Violier P. Le vieillissement de la population périurbaine – Quelles stratégies pour pallier la disparition d'une mobilité autonome ?. *Espace Populations Sociétés*. 2010/1, p. 83-93. <https://doi.org/10.4000/eps.3940>
- Pochet P., Corget R. Entre « automobilité », proximité et sédentarité. *Espace Populations Sociétés*. 2010/1, p. 69-81. <https://doi.org/10.4000/eps.4604>
- Rosson, M.B., Carroll, J. *Usability engineering: scenario- based development of human-computer interaction*. San Francisco, CA: Morgan Kaufmann Publishers Inc.
- Shambour Q.Y., Hussein A.H., Kharma Q.M., Abualhaj M.M. Effective hybrid content-based collaborative filtering approach for requirements engineering. *Computer Systems Science and Engineering*. 40(1), 2022: 113-125. Doi:10.32604/CSSE.2022.017221.
- Sokovic M., Pavletic D., Kern Pipan K. Quality improvement methodologies – PDCA cycle, RADAR matrix, DMAIC and DFSS. *Journal of Achievements in Materials and Manufacturing Engineering* (43:1), pp. 476-483, 2010.
- Touzani M., Ponsard C., Laurent A., Libourel T., Quinqueton J. Vers une modélisation et une analyse des exigences spatio-temporelles. *Inforsid 2016*, p. 51-66.
- Van Lamsweerde A. Goal-oriented requirements engineering: a guided tour. *In Int. Symposium on Requirements Engineering*. Toronto.Canada. IEEE, 2001, p.249-262.
- Verdier C., A. Front, D. Rieu, F. Forest. Innovations organisationnelles dans la prise en charge à domicile des personnes fragiles. *Revue Innovatio. Numéro 5 : L'interdisciplinarité en action au sein des projets de recherche en innovation*. Disponible sur : <https://innovatio.univ-grenoble-alpes.fr/index1e87.html?id=407>, 2018.
- Verdier C. *et al. Mobilité des personnes âgées - Etat de l'art*. Livrable n°1. Pack Ambition Recherche, 2019.
- Wanderley F., Belloir N., Bruel J.M., Hameurlain N, Araújo J. Des buts à la modélisation système : une approche de modélisation des exigences centrée utilisateur. *Inforsid 2014* : 113-128.

---

# Lambda+ Architecture et vérification de la conservation des propriétés avec la théorie des catégories

Annabelle Gillet, Éric Leclercq, Nadine Cullot

LIB Univ. Bourgogne Franche Comté, Dijon, France  
{prenom}.{nom}@u-bourgogne.fr

---

*RÉSUMÉ. La Lambda Architecture a été beaucoup critiquée, principalement à cause de sa complexité mais aussi car les propriétés de temps réel et d'exactitude des traitements sont effectives chacune dans une couche distincte mais pas dans l'architecture entière. Toutefois, la Lambda présente des mécanismes intéressants. Nous proposons donc une évolution de la Lambda Architecture : la Lambda+ Architecture, qui supporte à la fois des analyses exploratoires et en temps réel sur les données. Nous proposons également d'étudier la conservation des propriétés dans les compositions de composants d'une architecture avec la théorie des catégories.*

*ABSTRACT. The Lambda Architecture has been highly criticized, mostly because of its complexity and because the real-time and correctness properties are each effective in a different layer but not in the overall architecture. Nevertheless, it proposes some interesting mechanisms. We present a renewal of the Lambda Architecture: the Lambda+ Architecture, supporting both exploratory and real-time analyzes on data. We propose to study the conservation of properties in composition of components in an architecture using the category theory.*

*MOTS-CLÉS: Patron d'architecture, Théorie des catégories, Lambda Architecture*

*KEYWORDS: Architecture pattern, Category theory, Lambda Architecture*

---

Les propriétés d'exactitude des traitements, de temps réel et de tolérance aux pannes ont toujours été un enjeu majeur lors de la conception d'architectures. (Lampson, 1983) propose quelques suggestions qui sont toujours d'actualité et qui peuvent être retrouvées, entre autres, dans la Lambda Architecture (Marz, 2011). La couche *speed* de la Lambda permet d'obtenir la propriété de temps réel, tandis que la couche *batch* supporte la propriété d'exactitude des traitements. Mais lorsque les données sont rassemblées grâce à la couche *servng*, ces deux propriétés ne peuvent être supportées simultanément. Nous proposons d'étudier la conservation des propriétés dans les compositions de composants d'une architecture grâce à la théorie des catégories, mais aussi de faire évoluer la Lambda Architecture en la Lambda+ Architecture afin d'améliorer son support des propriétés et d'étendre ses cas d'utilisation pour pouvoir tirer profit des données massives. Cet article présente un résumé de l'article (Gillet *et al.*, 2021) publié dans la conférence CAiSE 2021.

La Lambda+ Architecture a deux fonctionnalités principales : 1) stocker les données d'une façon qui permet de réaliser des analyses exploratoires ; et 2) calculer en temps réel des indicateurs macroscopiques prédéfinis afin de suivre l'évolution de données d'intérêt. Pour cela, la Lambda+ est constituée de cinq composants interagissant de manière asynchrone à l'aide de messages. La dualité entre les analyses exploratoires et les indicateurs macroscopiques a une grande importance dans un contexte de données massives, dans lequel la combinaison du volume et de la variété des données empêche de révéler l'intégralité de leur valeur.

Dans le domaine de la recherche de la conception d'architectures logicielles, le développement d'une théorie adaptée accompagnée d'une formalisation est essentielle (Broy, 2011 ; Johnson *et al.*, 2012). Il est nécessaire de pouvoir prouver le maintien des propriétés dans l'ensemble d'une architecture, que ce soit au moment de sa conception ou de son évolution. La théorie des catégories (Eilenberg, MacLane, 1945) est une approche prometteuse pour répondre à ces besoins. En se concentrant sur les relations (les morphismes) et les compositions, il est possible de combiner de puissants mécanismes qui peuvent être ensuite appliqués aux architectures. De cette manière, le comportement des foncteurs couplés aux préordres permet d'étudier la conservation ou la perte des propriétés dans les compositions de composants. La formalisation que nous avons développée à partir de ces éléments a été appliquée à la Lambda Architecture ainsi qu'à la Lambda+ Architecture afin de montrer les défauts de la Lambda et la manière dont la Lambda+ les compense.

#### Remerciements

*Ce travail est soutenu par ISITE-BFC (ANR-15-IDEX-0003), piloté par Gilles Brachotte, laboratoire CIMEOS EA-4177, Université de Bourgogne.*

#### Bibliographie

- Broy M. (2011). Can practitioners neglect theory and theoreticians neglect practice? *Computer*, vol. 44, n° 10, p. 19–24.
- Eilenberg S., MacLane S. (1945). General theory of natural equivalences. *Transactions of the American Mathematical Society*, vol. 58, n° 2, p. 231–294.
- Gillet A., Leclercq É., Cullot N. (2021). Lambda+, the renewal of the lambda architecture: Category theory to the rescue. In *International Conference on Advanced Information Systems Engineering*, p. 381–396.
- Johnson P., Ekstedt M., Jacobson I. (2012). Where's the theory for software engineering? *IEEE software*, vol. 29, n° 5, p. 96–96.
- Lampson B. W. (1983). Hints for computer system design. In *Proceedings of the ninth acm symposium on operating systems principles*, p. 33–48.
- Marz N. (2011). *How to beat the cap theorem*. Consulté sur <http://nathanmarz.com/blog/how-to-beat-the-cap-theorem.html>

---

# Les exigences pour un choix d'architecture dans le cadre d'une migration dans les nuages

**Antoine Aubé<sup>1</sup>, Thomas Polacsek<sup>2</sup>, Clément Duffau<sup>3</sup>**

1. *Office National d'Études et de Recherches Aérospatiales*  
Toulouse, France

*antoine.aube@onera.fr*

2. *Office National d'Études et de Recherches Aérospatiales*  
Toulouse, France

*thomas.polacsek@onera.fr*

3. *Stack Labs*

Toulouse, France

*clement.duffau@stack-labs.com*

---

**RÉSUMÉ.** La migration d'un système vers un nuage est un ensemble d'activités permettant de tirer profit de l'informatique en nuage. Elle nécessite la conception d'un environnement d'exécution déployé par la configuration de services infonuagiques. Du choix de ces services et de leur configuration dépend la satisfaction de certaines exigences qui concernent le développement et l'exploitation du système à migrer. À travers une série d'entretiens semi-dirigés auprès de spécialistes de l'informatique en nuage, nous avons identifié dix exigences de haut niveau pour la sélection d'un environnement infonuagique lors d'une migration et leurs conséquences sur la conception de systèmes, ainsi qu'une classification de ces exigences.

**ABSTRACT.** A cloud migration is a set of activities to take advantage of cloud computing. It requires the design of an execution environment deployed by configuring cloud services. The satisfaction of certain development and operations-related requirements depends on the choice of these services and their configuration. Through a series of semi-structured interviews with cloud computing specialists, we identified ten high-level requirements for selecting a cloud environment during a migration and their implications for system design, together with a classification of these requirements.

**MOTS-CLÉS :** *Informatique en nuage, Migration, Ingénierie des exigences*

**KEYWORDS:** *Cloud Computing, Cloud migration, Requirements engineering*

---

## 1. Introduction

L'informatique en nuage, dit *Cloud Computing* en anglais, est un mode d'accès à des ressources informatiques comme des serveurs, du stockage, du réseau voire du logiciel, sous la forme de services accessibles via l'Internet. Une telle approche repose sur une infrastructure matérielle mise en commun pour tous les utilisateurs qui peuvent individuellement allouer et désallouer très rapidement des ressources suivant leurs besoins (Mell, Grance, 2011). Ainsi, en choisissant et en configurant ces ressources, l'utilisateur constitue un environnement sur lequel il peut déployer son application. Avec l'informatique en nuage, la notion d'infrastructure tend à s'effacer. Dans les faits, elle est masquée par un catalogue de services nommé un nuage. Les catégories communément utilisées pour classer ces services définissent les responsabilités portées par le fournisseur infonuagique et ses utilisateurs. Ces catégories sont, ordonnées par responsabilité croissante du fournisseur infonuagique : *IaaS*<sup>1</sup>, *PaaS*<sup>2</sup> et *SaaS*<sup>3</sup>.

Pour une entreprise, il peut être intéressant en termes de coûts de passer à l'informatique en nuage. En effet, de nombreuses entreprises proposent aujourd'hui des services infonuagiques, ce qui permet aux clients d'uniquement payer les services à l'usage, c'est-à-dire ce qu'ils consomment réellement. Dès lors, l'entreprise peut espérer réaliser des économies, puisqu'elle n'a plus à constituer et entretenir un parc de machines avec tous les coûts inhérents tels que le loyer, l'électricité, l'accès à l'Internet, etc. De plus, sans machine, elle s'exonère des tâches de maintenance et du personnel nécessaire à la maintenance.

Cependant, la réalité n'est pas si simple. On ne conçoit pas et on ne développe pas des systèmes infonuagiques comme on conçoit un système d'information classique. En effet, pour qu'un système infonuagique soit efficace, celui-ci doit utiliser les services adéquats, sans surconsommation tout en garantissant son exécution opérationnelle. Nous avons ici un double objectif qui, d'une part, est de définir et provisionner les bons services pour répondre au besoin du système et, d'autre part, de ne pas surprovisionner pour éviter des coûts inutiles.

Cet objectif est très difficile à gérer dans les phases en amont d'un cycle de développement. En effet, il n'est pas simple d'évaluer, en phase de spécification, les effets de l'utilisation de service sur des exigences portant, par exemple, sur les coûts ou la qualité du système. Dans la pratique, ce manque d'anticipation est la cause de dépassement de budget et de systèmes qui ne répondent pas tout à fait à leurs

---

<sup>1</sup>1. *IaaS* pour *Infrastructure as a Service*, service d'infrastructure à la demande : le service déploie des unités de calcul, du stockage et de la mise en réseau.

<sup>2</sup>2. *PaaS* pour *Platform as a Service*, service de plateforme à la demande : le service déploie un cadre d'exécution des logiciels et prend en charge la gestion du système d'exploitation sous-jacent. Dans la pratique, le client a le contrôle sur l'application déployée, mais pas sur l'infrastructure virtuelle.

<sup>3</sup>3. *SaaS* pour *Software as a Service*, service de logiciel à la demande : le service déploie un logiciel prêt à l'usage (e.g. Gmail, Gitlab.com).



exigences. Il est donc nécessaire de disposer de la capacité à évaluer, le plus tôt possible, les conséquences des choix de services sur les exigences du système. Encore faut-il comprendre les motivations qui mènent aux choix de services et aux choix du fournisseur infonuagique. Dès lors, même si ces motivations sont multiples, il peut être intéressant de dresser un panorama des éléments qui impactent fortement le choix du fournisseur et des services. Par éléments, nous entendons des exigences de très haut niveau, des buts et des objectifs. Ces buts peuvent être fonctionnels, comme garantir une performance pour un nombre de requêtes donné, ou non-fonctionnels, comme disposer d'un hébergement dans une zone géographique bien précise.

Pour pouvoir évaluer et ainsi fournir une aide à la conception de systèmes infonuagiques, il nous semble important de cartographier l'ensemble des dimensions qui sont cruciales pour les concepteurs de tels systèmes. C'est suivant ces buts, ces dimensions, qu'il sera possible d'évaluer des choix d'architectures et de faire des compromis entre plusieurs solutions.

Dans cette optique, nous avons mené une première enquête, au travers d'entretiens avec des spécialistes de la migration d'applications en informatique en nuage. La migration vers un nuage consiste tout simplement à faire migrer des applications d'un système d'information local à l'entreprise vers une solution utilisant de l'informatique en nuage. Dès lors, une telle migration implique des changements d'architecture et des choix sur les services qui vont être utilisés. De plus vient le problème du choix des fournisseurs de ces services. Nous allons ici présenter les premiers résultats de cette enquête.

Dans la Section 2, nous présentons l'objectif de notre recherche et la méthodologie que nous suivons. La Section 3 décrit l'échantillon de participants à l'enquête et la Section 5, suivi d'une proposition de classification de ces exigences en Section 6. Enfin, nous présentons des perspectives pour des travaux futurs en Section 7 avant de conclure en Section 8.

## **2. Objectif, motivations et méthodologie**

Notre objectif est de comprendre ce qui motive les choix d'environnements infonuagiques dans une opération de migration. Par environnement, nous parlons des ressources déployées par la configuration de services infonuagiques. Nous avons choisi de nous limiter ici à la migration afin de ne pas étendre le champ des possibles à tous les projets de systèmes infonuagiques. Ce cadre permettant, nous l'espérons, de dégager un premier corpus restreint d'exigences. Nous pouvons ramener notre but à la question suivante :

*Quelles sont les principales exigences de la sélection d'un environnement dans le cadre de la migration vers un nuage ?*

Pour pouvoir répondre à cette question, nous avons décidé de mener des entretiens de spécialistes du domaine. Notre approche est motivée par le manque de données dans la littérature. En effet, si de nombreux travaux s'intéressent à identifier

des stratégies de choix d'environnement, le sujet des motivations des organisations qui dirigent une migration vers un nuage est peu adressé.

Citons, pour exemple, les travaux de Quinton (2014) qui propose une méthode basée sur les lignes de produits logiciels pour sélectionner un environnement à partir de l'expression des besoins du système. Dans le même esprit, Ferry *et al.* (2013) proposent un langage de modélisation d'infrastructure afin de faciliter l'utilisation de ressources infonuagiques. L'utilisateur exprime les besoins non fonctionnels de son système dans un modèle indépendant de tout fournisseur infonuagique, puis ce modèle est instancié pour un fournisseur donné et utilisé à l'exécution pour le déploiement, la surveillance et l'approvisionnement dynamique des ressources. Toujours dans une logique d'aide à la sélection de service, Zhang *et al.* (2012) ont conçu un système de recommandation pour trouver la meilleure configuration de machines virtuelles en fonction d'exigences non fonctionnelles. Leur approche se base sur l'ontologie des ressources de l'informatique en nuage : *Cloud Computing Ontology* (CoCoOn). Pour finir, Amiri *et al.* (2017) proposent d'utiliser l'apprentissage automatique pour optimiser le provisionnement des ressources.

Comme nous le voyons, tous ces travaux adressent la satisfaction des exigences, mais ne se focalisent pas sur leur identification ni sur les motivations sous-jacentes.

Forts de ce constat, nous avons opté pour une approche basée sur un retour d'expérience. Plus précisément, nous avons mené une série d'entretiens individuels qualitatifs semi-structurés auprès de professionnels. L'entretien qualitatif est une pratique courante et assez ancienne dans le champ de l'ethnographie et de la recherche médicale (DiCicco-Bloom, Crabtree, 2006). Il permet au travers d'une conversation approfondie, suivant un protocole préétabli, d'acquérir une meilleure compréhension du sujet d'étude et de formuler des hypothèses. Généralement semi-structurés, de tels entretiens consistent à poser des questions ouvertes prédéterminées.

Notre méthodologie a été la suivante. Nous avons mené une série d'entretiens individuels. Ces entretiens ont été menés en face à face, ou en visioconférence, sur une durée allant de 45 minutes à une heure. Pour faciliter la retranscription, ils ont été enregistrés avec l'accord des participants.

Les entretiens ont toujours suivi le même format avec, pour commencer, une phase d'introduction avec une explication du sujet de recherche, une explication du périmètre de la discussion afin de limiter les digressions sur des sujets connexes, et une présentation du déroulement de l'entretien.

L'entretien étant semi-structuré nous avons défini un jeu de questions préétablies posées systématiquement et toujours dans le même ordre. La première série de questions se rapporte au participant, ses rôles au sein de l'équipe, et au contexte de la migration (taille de l'équipe de développement, domaine de l'entreprise, etc.). Ensuite commencent réellement les questions concernant notre sujet d'étude. Nous avons dérivé notre question de recherche sous la forme des dix questions ouvertes ci-dessous :

1. Comment avez-vous analysé les composants à migrer ? Quelles étaient les informations importantes à en tirer ?
2. Quelle a été la démarche d'analyse pour trouver les besoins spécifiques à l'informatique à nuage ? Autrement dit, pour identifier les exigences concernant la sécurité, la qualité de service, etc.
3. Sous quelle forme avez-vous pu consulter les exigences auxquelles répond le système ?
4. Avez-vous des preuves que le système existant répondait effectivement aux exigences ?
5. La façon dont l'utilisateur se sert du système était-elle un élément important à capturer à cette étape ? Si oui, comment cela a-t-il été fait ?
6. Avez-vous des exigences sur la qualité de service ? Si oui, quels aspects de la qualité de service ?
7. Quand on parle budget opérationnel, est-ce que le coût de la main-d'œuvre est pris en compte ? Et si oui, comment est-il calculé ?
8. Avez-vous des exigences sur la performance environnementale du système ?
9. Pour le budget opérationnel, comment sont exprimées les exigences ? Est-ce un budget global au système, un par unité fonctionnelle, par ressource, par volumétrie d'entrée, etc. ? Est-ce que ce sont des sommes fixes ou bien est-ce plus complexe que ça ? Est-ce que nous parlons de seuils de tolérance, de marge, etc. ?
10. L'effort d'adaptation du code a-t-il été une contrainte ?

### 3. Profils des interlocuteurs

Afin d'obtenir des réponses pertinentes lors des entretiens, McCracken (1988) recommande la constitution d'un échantillon homogène ayant une proximité avec la question de recherche. Pour cette raison, nous avons choisi de nous focaliser sur des personnes proches des aspects techniques d'une migration vers un nuage et qui sont intervenus dans les mêmes conditions. Pour mener notre enquête, nous avons interrogé sept professionnels qui travaillent dans des sociétés de services numériques extérieures aux organisations qui possèdent le système à migrer. Concernant la proximité avec notre objet d'étude, tous ces professionnels occupent un poste en lien avec le choix de l'architecture de l'environnement infonuagique, que cela soit en terme de conseils, d'étude ou de décisions techniques. Ainsi nous avons interrogé un responsable projets, trois architectes de systèmes infonuagiques, deux spécialistes de l'exploitation opérationnelle et un expert de la migration, qui a plus d'une cinquantaine de migrations à son actif.

Sans prétendre à l'exhaustivité, nous avons cherché des personnes qui ont participé à des migrations très différentes, aussi bien dans la taille des projets et leurs objectifs que dans leurs contextes. En termes de contexte, les projets couverts par nos entretiens adressaient : la migration de systèmes pour une entreprise qui gère

de très gros volumes de données avec des clients dans le monde entier, des migrations pour un établissement public, une migration pour un service à destination des collectivités locales et des migrations de plusieurs centaines de systèmes pour un grand groupe de distribution. Concernant la taille des projets, certains systèmes étaient maintenus par une équipe très réduite avec un développeur commun à plusieurs systèmes qui intervient également sur l'exploitation des systèmes, un ou deux experts du métier qui sont aussi développeurs et un chef de projet. De l'autre côté du spectre, nous avons aussi pu discuter de migrations de systèmes maintenus par des équipes de plus de cinquante personnes.

#### 4. Entretiens

Une première remarque préliminaire s'impose concernant le contenu des entretiens. À notre grande surprise, il est apparu que les développeurs ne connaissaient pas le terme *exigence* et que la définition d'exigences n'était pas toujours exsangue de certaines ambiguïtés pour les architectes. Cependant, au vu des entretiens, nous sommes bien dans l'identification des objectifs, des buts et des contraintes du client qui mènent à définir l'environnement infonuagique, nous avons bien des interlocuteurs qui nous parlent d'exigences de haut niveau. Par conséquent, dans un souci de clarté, nous utiliserons le mot « *exigence* » dans la suite, même si ce mot n'a pas été employé par les personnes interrogées.

Par ailleurs, lors des entretiens, il est important de noter que nos différents interlocuteurs ont relaté des difficultés à accéder à certaines exigences. Par exemple, dans le cadre d'un projet consistant à identifier les problèmes de systèmes existants et à proposer un environnement infonuagique pour une migration, il a été fait état d'une non-connaissance du budget opérationnel du futur système : « *Je pense qu'ils l'avaient en interne, mais à aucun moment ils nous l'ont dit ; je pense qu'ils attendaient [notre proposition] pour voir si c'était une solution qui irait pour leur besoin.* ». Sur ce point, nous ne pouvons faire que des supputations. Il est possible que l'information concernant le budget n'ait pas été communiquée dû à la relation commerciale entre l'entreprise et l'intervenant, ou parce que l'entreprise, manquant de connaissance sur l'informatique en nuage, n'était pas en mesure d'estimer ce qu'elle était prête à payer pour l'opération du système.

Un élément important, présent dans tous nos entretiens et dans tous les projets auxquels ont participé tous nos interlocuteurs, est la présence d'exigences liées aux coûts. Ceci n'est pas étonnant. En effet, la réduction des coûts est la promesse faite aux organisations pour promouvoir l'informatique en nuage. Dès lors, il est normal que les coûts soient un aspect fondamental dans la sélection de l'environnement. Nous devons toutefois faire ici le tri parmi les coûts possibles. Premièrement, nous avons les coûts liés aux activités de la migration. Ces coûts adressent tous les développements logiciels éventuellement réalisés pour adapter le système ainsi que toutes les opérations de déploiement sur le nuage. Deuxièmement, les coûts opérationnels. Ce sont des coûts récurrents et ils concernent le l'exploitation du système.

Concernant les coûts opérationnels, nous pouvons aussi les diviser en deux catégories : les coûts de l'environnement infonuagique et les coûts liés à la maintenance opérationnelle. Les coûts de l'environnement infonuagique correspondent tout simplement à ce que facture le fournisseur infonuagique par l'utilisation de ses services. Ils sont parfois estimés à l'avance grâce à des outils édités par ces fournisseurs. Les coûts de maintenance opérationnelle correspondent, eux, au personnel de l'organisation chargé d'exploiter le système, et ils sont très difficiles à anticiper.

## 5. Exigences motivant les choix d'architecture

À partir des entretiens, nous avons identifié dix catégories d'exigences, d'exigences de haut niveau, motivant les choix lors de la constitution d'un environnement infonuagique dans le cadre d'une migration. Elles sont présentées dans cette section.

Notons que ces exigences ne sont pas apparues nommément dans le discours des personnes interrogées. Leur identification relève de notre part d'une généralisation d'un ensemble d'anecdotes et d'exemples d'exigences et de contraintes très factuelles données par les personnes interrogées. Par exemple, une anecdote collectée relevait d'une contrainte sur l'utilisation d'un langage de programmation : *« Il fallait que ça soit maintenable par eux en termes de développement. [...] Par exemple, les scripts en Python, on ne pouvait pas proposer qu'ils soient réécrits en Go, car ils ont des ingénieurs qui viennent du monde de l'intelligence artificielle, qui font du Python, car c'est très répandu dans leur domaine, et c'est difficile de leur faire changer de langage du jour au lendemain »*. En prenant du recul, nous constatons qu'il s'agit d'un cas particulier d'un critère plus général : minimiser le besoin de nouvelles compétences pour développer et opérer le système.

Notre étude étant limitée à un échantillon très restreint, il serait illusoire de chercher à déterminer une quelconque relation de préférence entre ces exigences. Dans le cadre de cet article, nous avons donc décidé de présenter ces dix exigences en les classant en ordre décroissant suivant le nombre d'interlocuteurs qui les ont évoquées.

### 5.1. Minimiser la maintenance opérationnelle

La maintenance opérationnelle est l'effort fourni par des opérateurs humains pour entretenir le bon fonctionnement du système et assurer la livraison dans de bonnes conditions du service.

La diminution de cet effort est apparue dans l'ensemble de nos entretiens comme un critère de sélection de services infonuagiques. Au vu des entretiens, il apparaît clairement que le but sous-jacent à cette exigence est la diminution des coûts de la masse salariale dévolue à l'exploitation du système. Selon le contexte, elle ne s'instancie pas de la même façon, car, dans certains projets, elle vise à réduire la

charge de travail des équipes déjà en place là où, dans d'autres, c'est en prévision du futur pour limiter l'embauche de nouveaux spécialistes.

Comme nous l'a rapporté un de nos interlocuteurs, cette exigence peut amener à une adaptation de l'architecture de migration. Dans son cas, il citait l'exemple de l'automatisation de tâches manuelles en utilisant un service de *Apache Airflow*<sup>4</sup> à la demande.

Dans le cadre de projets visant à proposer des environnements infonuagiques pour des systèmes de traitement de très grosses données, les architectures privilégiaient systématiquement l'utilisation de services *serverless*<sup>5</sup> afin de réduire significativement le besoin d'interventions humaines pour assurer la continuité du service.

Dans le cadre de la migration de plusieurs applications impliquant des équipes de maintenance de l'ordre d'une cinquantaine de personnes, tous les systèmes ont été migrés vers un *PaaS* car la maintenance des plateformes monopolisait beaucoup les équipes opérationnelles.

## 5.2. Minimiser les coûts de l'environnement

Comme évoqué précédemment, un environnement infonuagique est déployé par des services dont les coûts dépendent de l'usage fait du système.

Pour accomplir un même rôle (e.g. stocker des fichiers, exécuter un programme), plusieurs services sont candidats et ne facturent pas le même usage. Par exemple, pour le stockage de fichiers, certains services reposent sur une machine virtuelle dont le coût est facturé (i.e. le temps de CPU et RAM, licence de système d'exploitation, etc.) avec un supplément pour l'infogérance ; d'autres facturent uniquement les interactions (i.e. lectures, écritures, etc.) et le volume de données moyen stocké pendant la période facturée. Le choix de la solution la plus économe est en fait dépendant du modèle d'utilisation du système.

Pour un projet d'hébergement d'une forge logicielle (i.e. une plateforme en ligne qui agrège des outils à l'usage des développeurs pour sauvegarder les codes sources, effectuer des validations, gérer des projets, etc.), en outre des préoccupations quant à la performance, la personne interrogée a décidé d'opter pour un *IaaS* pour pouvoir configurer une instance correspondant le plus possible aux besoins du système. Les *PaaS* qu'il aurait pu envisager lui permettait de configurer des instances soit insuffisantes, soit surdimensionnées. En sélectionnant plus finement la configuration de l'instance, il a ainsi réduit les coûts de l'environnement au détriment de l'effort de maintenance opérationnelle.

Tous les participants ont indiqué que la diminution des coûts de leur environnement était un critère de sélection des services, mais seul un en a fait une priorité. Dans la plupart des cas, cette diminution des coûts de l'environnement

<sup>4</sup>4. Logiciel d'orchestration de tâches de calcul. <https://airflow.apache.org/>

<sup>5</sup>5. Services dans lesquels le fournisseur de services gère entièrement les ressources allouées.

semble moins importante que d'autres critères. Par exemple, dans le cas du projet utilisant *Apache Airflow*, il aurait été envisageable d'employer des services *serverless* coûtant moins cher d'après leur modèle d'utilisation du système, mais cela aurait nécessité des connaissances que n'a pas l'équipe de développement.

### 5.3. Maximiser la fiabilité du système

Dans le cas d'un hébergement de l'infrastructure informatique sur site, remplacer des ressources peut prendre du temps (e.g. achat, transport, réception, mise en place du matériel) et se protéger d'un incident est coûteux (e.g. réplication des données dans plusieurs centres informatiques suffisamment éloignés).

Quatre de nos interlocuteurs sont intervenus sur des projets pour lesquels la migration a été initiée suite à un tel incident. Ces derniers peuvent être liés, par exemple, à des aléas climatiques, à des actions malveillantes ou à l'usure du matériel, qui ont occasionné des coupures du réseau, la destruction de disques durs ou d'unités de calcul. À cause de ces incidents, leur système n'a pas été utilisable pendant un temps et certaines données ont été perdues.

L'un des objectifs de ces migrations était donc l'amélioration de la fiabilité de ces systèmes.

Un des architectes interrogés rapporte : « *Le but [de la migration] pour l'exploitation de leur système, c'était d'apporter de la robustesse [...] Le but premier, c'était de fiabiliser, car, avant, tout reposait sur quelques machines et un ingénieur qui opérait l'ensemble. [...] Avec ce qu'ils avaient avant, s'il y avait une grosse panne, ils auraient eu beaucoup de mal à tout remettre en ordre* ».

Pour ce faire, les risques ont été identifiés et un Plan de Reprise d'Activité (*PRA*) a été mis en place. Ce *PRA* décrit les procédures permettant de réagir à une panne, et repose sur l'utilisation de services qui fournissent des garanties de fonctionnement (*Accord de niveau de service*<sup>6</sup>) par exemple pour la durabilité et la disponibilité des données, ainsi que sur l'automatisation de la réplication des ressources (e.g. données, machines virtuelles) sur plusieurs zones géographiques.

### 5.4. Maximiser les performances

La performance est un critère que l'on retrouve dans de nombreux projets, mais pas dans tous et ils n'adressent pas tous la même notion de performance.

Plus précisément, dans le contexte d'une entreprise dont l'infrastructure était hébergée en France, l'arrivée d'un nouveau client basé en Amérique du Sud a introduit une exigence de réduction de la latence des services spécifiquement pour ce client. Le choix du fournisseur infonuagique a donc été fait en tenant compte de

---

<sup>6</sup> De l'anglais *Service-Level Agreement*, engagement du fournisseur de services sur la qualité du service qu'il délivre.

la possibilité de déployer des ressources à la fois en Europe de l'Ouest et en Amérique du Sud.

Un des objectifs de la migration de la forge logicielle évoquée précédemment était de diminuer les temps d'exécution des tâches de calculs intensifs (e.g. compilation des logiciels). Cette exigence a motivé l'utilisation de services de machines physiques à la demande (sans surcouche de virtualisation coûtant du temps de calcul) dotées de GPU. Cette solution a été préférée par rapport à l'utilisation d'un *SaaS* ou d'un hébergement sur un *PaaS*, qui auraient été des solutions plus simples à gérer. À ce sujet, la personne interrogée nous a dit : « *Il y avait deux choix : soit partir sur du Kubernetes infogéré, soit monter soi-même son cluster Kubernetes. On a fait le second choix pour des raisons de performances. C'est-à-dire que sur les machines virtuelles que propose [le fournisseur infonuagique] derrière leur Kubernetes infogéré, on ne retrouvait pas les caractéristiques dont on avait besoin, ou en tout cas ça aurait eu un coût trop important [...]* ».

### **5.5. Minimiser le besoin de nouvelles expertises**

Travailler avec des services infonuagiques requiert l'acquisition de connaissances techniques très spécifiques (e.g. limitations, protocoles) aussi bien pour le développement que pour l'exploitation. Ces connaissances sont plus ou moins nombreuses suivant les services utilisés. Par exemple, l'administration d'une machine virtuelle est proche de l'administration d'une machine réelle, mais n'est pas comparable avec l'administration d'une ressource *serverless*.

Or, certains projets qui nous ont été présentés ont une équipe de développement réduite avec, par exemple, des développeurs chargés également d'administrer le système. Dans aucun de ces projets, l'entreprise n'envisageait d'agrandir l'équipe, pour des raisons de coûts, ni de remplacer du personnel, par crainte de perdre une expertise métier propre à l'organisation. Le choix des services devait donc minimiser le nombre de nouvelles connaissances à acquérir pour les équipes en place.

Précédemment, nous avons évoqué un retour au sujet d'une migration dans laquelle le langage de développement devait absolument rester *Python*. Cette contrainte a eu de véritables conséquences sur le choix des services infonuagiques : la plupart des *PaaS* prennent en charge en ensemble limité de langages.

Concernant le choix du fournisseur infonuagique, dans le cadre de la migration d'un *SaaS*, un de nos interlocuteurs évoque le fait que l'équipe qui administre le système avait de l'expérience avec un fournisseur et que cela a grandement encouragé le choix de rester avec celui-ci au détriment de services proposés par d'autres.



### 5.6. *Minimiser les coûts de migration*

Pour bénéficier des avantages de l'informatique en nuage (i.e. élasticité, etc.) ou pour optimiser les coûts de l'environnement infonuagique, il est nécessaire d'adapter les applications. Ces modifications correspondent à du développement logiciel et peuvent considérablement augmenter les coûts de la migration.

Cependant, dans certains projets avec un budget réduit alloué à la migration, il a été décidé de n'apporter que le strict minimum de modifications aux applications, ce qui apporte une contrainte au choix des services capables de les héberger.

En effet, dans certains cas spécifiques, un composant logiciel peut être hébergé sur un *PaaS* sans changement de son code s'il respecte certaines contraintes, comme a remonté l'un des architectes pour un logiciel écrit en Java et exécuté sur un serveur Tomcat qui a été migré sur *Google App Engine*<sup>7</sup>. L'alternative est d'héberger ces applications sur des machines virtuelles, qui permettent une plus grande liberté au prix d'une facturation moins intéressante et d'une moindre infogérance du fournisseur infonuagique.

### 5.7. *Minimiser la durée de migration*

Certaines migrations sont contraintes par leur durée, avec une échéance de fin fixée en amont qui ne peut pas être changée.

Ainsi, l'un des participants, qui est intervenu dans un projet pour l'agriculture, nous a rapporté « *Comme leur activité est saisonnière, il fallait finir la migration avant la saison qui arrivait* ». Pour tenir les délais, ils n'ont pas effectué un changement de type de base de données qui aurait permis de limiter le coût de l'environnement infonuagique. Également, dans le cadre de la migration d'application de gestion de données massives pour un service international, eu égard au temps alloué pour effectuer la migration, il a été décidé de réutiliser des morceaux de logiciels existant dans le but de ne pas investir de temps et de risque de dérive dans des opérations d'adaptation du système.

Nous sommes donc bien face à une exigence différente de la minimisation du budget de migration, car ici, c'est la date de mise en production qui a dirigé les choix d'architecture.

### 5.8. *Garantir la souveraineté numérique*

Le concept de souveraineté numérique, qui ne renvoie pas à des contraintes ni des pratiques précises, est de plus en plus prégnant dans le monde de l'informatique en nuage. Des réglementations telles que le *CLOUD Act* pour les États-Unis ou le Règlement Général sur la Protection des Données (*RGPD*) pour l'Union Européenne ne viennent que renforcer cette thématique.

---

<sup>77</sup>. App Engine, un PaaS de Google Cloud. <https://cloud.google.com/appengine>

Dans la pratique, certaines organisations, en particulier celles qui manipulent des données sensibles pour un État, tendent à prendre en compte localisation des données et des traitements tout comme la législation qui s'applique aux fournisseurs infonuagiques.

Ainsi, cette exigence est apparue lors de projets de migration de deux systèmes pour un établissement public et lors d'une migration d'un *SaaS* à destination des collectivités locales. Concrètement, il était demandé que les données et traitements soient sur le territoire français et que le fournisseur de services soit sous législation française. D'après l'un des spécialistes de l'exploitation opérationnelle : « *Le DSI de [l'entreprise] souhaitait que la solution reste en France. C'était pour du nucléaire français, il n'y avait pas vraiment de raison que ça aille chez [un nuage états-unien] [...]* ». Selon le contexte, il semble que la mise en œuvre satisfaisante de cette exigence puisse prendre des formes variées, car l'expert en migrations a rapporté : « *Ils voulaient que leurs données soient sur un cloud souverain, on leur a proposé un entre-deux : mettre leurs données sur [un nuage états-unien] et la clé de chiffrement sur un cloud français* ».

### **5.9. Maximiser la sécurité**

De façon surprenante, une seule des personnes interrogées a évoqué des exigences liées à la sécurité. Dans son cas, suite à une première étape de migration sur un *IaaS*, des modifications ont été apportées et certains services ont été proscrits. En effet, l'utilisation de ces services impliquait de facto une modification de la topologie du réseau. Cette modification était vue comme un risque de sécurité dans ce contexte. Le système héberge des logiciels que l'État français impose de protéger de tout risque d'espionnage dans le cadre de la *Protection du potentiel scientifique et technique de la Nation*.

De façon générale, les exigences de sécurité peuvent être un frein à l'utilisation de certains services et configurations de services.

### **5.10. Maximiser l'agnosticisme à un fournisseur infonuagique**

L'agnosticisme à un fournisseur infonuagique d'un système peut être recherché pour diverses raisons, comme pour simplifier le développement de composants logiciels pour plusieurs plateformes (e.g. qui doivent être exécutés aussi bien sur site que sur divers fournisseurs infonuagiques), ou bien pour simplifier les migrations depuis un nuage-hôte si les conditions d'hébergement venaient à changer et ne plus être compatibles avec les exigences du système (e.g. tarifs, politiques).

Une manifestation de cette exigence est la volonté d'utiliser des logiciels libres.

Ainsi, dans le cadre d'une migration d'applications relevant de la sûreté de l'État, l'un de nos interlocuteurs a fait savoir la nécessité d'utiliser des logiciels libres afin d'éviter tout enfermement propriétaire, que cela concerne le fournisseur infonuagique comme la société prestataire qui administre le système. La crainte était

d'être, in fine, prisonnier des tarifs d'une entreprise. Pour cette raison, un service de *Kubernetes* à la demande a été choisi pour héberger leurs portails web plutôt qu'un *PaaS* propriétaire : « *On a choisi Kubernetes, car c'est un standard industriel libre [...] pour que l'entreprise soit le plus indépendant possible vis-à-vis d'autres entreprises* ».

## 6. Propositions de classification des exigences

Suite à la mise en évidence de ces dix exigences de haut niveau, nous avons cherché à établir une classification. Classiquement, l'ingénierie des exigences distingue les exigences fonctionnelles, qui décrivent les fonctions du système à développer, et non-fonctionnelles, qui décrivent les qualités désirées du système (Kotonya, Sommerville, 1998). En général, une migration ne transforme pas le produit, mais son support. Dès lors, la plupart des exigences que nous avons identifiées touchent peu au fonctionnel, cette classification n'est pas ici pertinente.

De leur côté, les approches orientées buts distinguent les *buts durs* qui peuvent être satisfaits sans ambiguïté, et des *buts mous* (*soft-goal*), ou *qualités*, qui décrivent une direction sans critère de satisfaction clair (Dalpiaz *et al.*, 2016 ; Lamsweerde, 2001). D'après les entretiens menés, certaines exigences peuvent être classées aisément dans l'une de ces deux catégories. Par exemple, l'exigence de souveraineté numérique est un *but dur*, car elle fixe sans ambiguïté ce qui est acceptable pour la législation que suit un fournisseur infonuagique ou la localisation d'une donnée. De même, si tous nos interlocuteurs ont manifesté le besoin de minimiser la maintenance opérationnelle de leur système, aucun chiffre n'a été avancé ; cette exigence est donc un *but mou*. Cependant, certaines exigences, telles qu'énoncées dans les entretiens, peuvent être rangées à la fois comme un *but mou* ou *dur*. Par exemple, minimiser le temps de migration est, selon le contexte, un *but mou*, si l'on cherche seulement à minimiser le temps, ou un *but dur*, si la migration doit être terminée avant une date fixée. Par ailleurs, même si ce n'est pas apparu lors de nos entretiens, les exigences comme la minimisation des coûts ou l'amélioration des performances sont propices à être mesurées (e.g. "diminuer la latence réseau" contre "la latence réseau ne dépasse pas tel seuil") et donc leur classification dépendrait du contexte également. Pour ces raisons, nous ne considérerons pas cette classification.

Finalement, notre choix s'est porté sur la dichotomie : développement logiciel et exploitation du système (DevOps). Il nous faut cependant prendre en compte un troisième élément : le fournisseur infonuagique. En effet, de nombreuses exigences reposent sur la capacité du fournisseur infonuagique à délivrer les services adéquats. Par conséquent, nous proposons la classification suivante :

- Les exigences liées au **développement logiciel**. Les choix de services découlant de ces exigences vont interférer avec le développement (dans le cadre de la migration, il s'agit d'adaptations) des logiciels. Les exigences liées au développement logiciel sont : **maximiser l'agnosticisme à un fournisseur infonuagique, minimiser le besoin de nouvelles expertises, minimiser les coûts et la durée de migration**.

- Les exigences liées à l'**exploitation du système**. Les efforts investis dans l'exploitation du système dépendent des choix de services découlant de ces exigences. Les exigences liées à l'exploitation du système sont : **maximiser la fiabilité du système, minimiser le besoin de nouvelles expertises, minimiser la maintenance opérationnelle**.
- Les exigences liées au **fournisseur infonuagique**. Ces exigences sont satisfaites uniquement par des propriétés du fournisseur. Les exigences liées au fournisseur infonuagique sont : **garantir la souveraineté numérique, maximiser les performances, minimiser les coûts de l'environnement, maximiser la sécurité**.

## 7. Travaux futurs

Sur la base de l'analyse présentée dans cet article, nous avons commencé une enquête que nous diffusons plus largement auprès d'experts industriels. Nous attendons de cette enquête qu'elle nous permette de confirmer la pertinence des exigences que nous avons présentées ici et d'identifier de nouvelles exigences de haut niveau qui n'auraient pas été évoquées dans nos entretiens comme, par exemple, les exigences relatives aux impacts environnementaux des systèmes infonuagiques, qui sont pourtant de plus en plus souvent imposées dans les appels d'offres, en particulier pour les marchés publics.

Notre classification n'est pas la seule possible. Classiquement, dans la gestion de projets, on identifie trois principales contraintes : la maximisation de la qualité, la minimisation des délais et la minimisation des coûts. Il est peut-être possible de classer les exigences que nous présentons dans cet article comme des raffinements de ces contraintes. En effet, certaines exigences visent à minimiser les coûts (minimiser les coûts de migration, minimiser le besoin de nouvelles expertises, minimiser la maintenance opérationnelle), d'autres à minimiser les délais (minimiser la durée de migration, minimiser le besoin de nouvelles expertises) et enfin certaines à maximiser la qualité du système (maximiser la sécurité, maximiser l'agnosticisme à un fournisseur infonuagique, garantir la souveraineté numérique, maximiser les performances, maximiser la fiabilité du système). Des travaux futurs devraient explorer les classifications possibles et chercher les plus efficaces pour guider la gestion des projets de migration.

Nos dix exigences ne sont pas indépendantes. Par exemple, minimiser la maintenance opérationnelle implique des développements pour utiliser des services hautement infogérés et rentre donc en conflit avec l'exigence de minimiser la durée de migration. Afin d'assister la prise de décision dans le cadre d'un compromis entre les contraintes d'un projet de migration, il serait intéressant d'identifier les liens entre nos exigences de haut niveau. Une fois ces liens identifiés, il pourrait être intéressant de définir un cadre méthodologique, basé sur les approches orientées buts, pour déterminer les compromis acceptables, et ainsi mieux appréhender la complexité cognitive inhérente à une migration vers un nuage public.

Les *stratégies de migration* classifient les migrations selon les activités et les choix de conception de l'environnement. Si la littérature scientifique et industrielle propose plusieurs catégorisations de stratégie de migration, il manque une méthode holistique pour déterminer quelle stratégie employer dans une situation donnée. Nous retrouvons ce constat dans les travaux de Zhao et Zhou (2014) qui identifient des défis à traiter sur la thématique des stratégies de migration. Une telle méthode holistique devrait prendre en compte les exigences que nous avons identifiées ici.

Enfin, l'évaluation de la satisfaction de certaines de ces exigences dès la conception n'est pas triviale. C'est par exemple le cas de l'exigence "minimiser les coûts de l'environnement" qui requiert d'estimer le coût de l'environnement infonuagique, qui dépend de la charge utilisateur à tout instant. Dès lors, il est coûteux pour une organisation d'identifier les situations dans lesquelles il est nécessaire de faire des compromis, et de formuler des compromis acceptables quand c'est nécessaire. En outre, des travaux apportent des approches pour évaluer individuellement certaines qualités d'un environnement infonuagique, comme ceux de Gesvindr *et al.* (2017) pour la performance des *PaaS* ou Belli *et al.* (2016) pour l'estimation des coûts des *IaaS*, mais il manque d'approches permettant de comparer entre elles des architectures proposant les mêmes fonctionnalités tout en étant très différentes (e.g. une architecture basée sur des *IaaS* et une autre *serverless*).

## 8. Conclusion

La migration d'un système vers un nuage permet aux entreprises de bénéficier des avantages de l'informatique en nuage. Si la littérature décrit plusieurs méthodes pour satisfaire certaines exigences, elle fait fi des contraintes que ces dernières s'imposent entre elles. En outre, il manque de travaux sur l'identification des exigences que peut avoir l'industrie.

Dans cet article, nous avons présenté l'analyse de retours de professionnels de l'informatique en nuage au sujet de la migration de systèmes vers un nuage. Ces retours ont été collectés à travers une série d'entretiens semi-formels. Cette analyse a mis en évidence dix exigences de haut niveau qui dirigent le choix des services infonuagiques et de leur configuration lors de la conception, pour lesquelles nous avons proposé une classification.

## Bibliographie

- Amiri M., Khanli L. M. (2017). Survey on prediction models of applications for resources provisioning in cloud. *J. Netw. Comput. Appl.*, vol. 82, p. 93-113.
- Belli O., Loomis C., Abdennadher N. (2016). Towards a cost-optimized cloud application placement tool. *IEEE International Conference on Cloud Computing Technology and Science*, p. 43-50.
- Dalpiazz F., Franch X., Horkoff J. (2016). iStar 2.0 langage guide. *CoRR*, vol. abs/1605.07767.

- DiCicco-Bloom B., Crabtree B. F. (2006). The qualitative research interview. *Medical education*, vol. 40, n°4, p. 314-321.
- Ferry N., Rossini A., Chauvel F., Morin B., Solberg A. (2013). Towards model-driven provisioning, deployment, monitoring, and adaptation of multi-cloud systems. *2013 IEEE Sixth International Conference on Cloud Computing*, p. 887-894.
- Gesvindr D., Buhnova B., Gasior O. (2017). Quality evaluation of PaaS Cloud application design using generated prototypes. *IEEE International Conference on Software Architecture*, ICSA 2017, p. 31-40.
- Kotonya G., Sommerville I. (1998). *Requirements engineering – processes and techniques*. John Wiley & Sons.
- Lamsweerde A. van (2001). Goal-oriented requirements engineering: A guided tour. *5th IEEE International Symposium on requirements engineering*, p. 249.
- McCracken G. (1988). *The long interview*. SAGE Publications, Inc.
- Mell P., Grance T. (2011). *The NIST definition of Cloud Computing*. Rapport technique n°800-145.
- Quinton C. (2014). *Cloud Environment Selection and Configuration: a software product lines-based approach*. Thèse de doctorat, Université Lille I.
- Zhang M., Ranjan R., Nepal S., Menzel M., Haller A. (2012). A declarative recommender system for cloud infrastructure services selection. *Economics of grids, clouds, systems, and services – 9th International Conference*, vol. 7714, p. 102-113.
- Zhao J., Zhou J. (2014). Strategies and Methods for cloud migration. *Int. J. Autom. Comput.*, vol. 11, p. 143-152.

---

## Profilage de services pour la gestion de l'énergie dans les architectures orientées services

**Jorge Andrés Larracoechea<sup>1</sup>, Philippe Roose<sup>1</sup>, Sergio Ilarri<sup>2</sup>,  
Yudith Cardinale<sup>3</sup>, Sébastien Laborie<sup>1</sup>**

<sup>1</sup>LIUPPA/E2S, Université de Pau et des Pays de l'Adour, Anglet, France

<sup>2</sup>Instituto de Investigación en Ingeniería de Aragón /I3A, Universidad de Zaragoza, Zaragoza, Spain

<sup>3</sup>Dpto. de Computación y T.I, Universidad Simón Bolívar, Caracas, Venezuela

jorge-andres.larracoechea@etud.univ-pau.fr,  
{Philippe.Roose, Sebastien.Laborie}@iutbayonne.univ-pau.fr,  
silarri@unizar.es,  
ycardinale@usb.ve

---

REFERENCE DE L'ARTICLE INTERNATIONAL. *Cet article est un résumé de l'article présenté lors de la conférence WebIST dont la référence est la suivante : Jorge Andrés Larracoechea, Philippe Roose, Sergio Ilarri, Yudith Cardinale, Sébastien Laborie, Mauricio Jacobo González – Towards Services Profiling for Energy Management in Service-Oriented Architectures – pp. 209-216 - WEBIST (International Conference on Web Information Systems and Technologies) – October 26-28, 2021 – Best Student Paper*

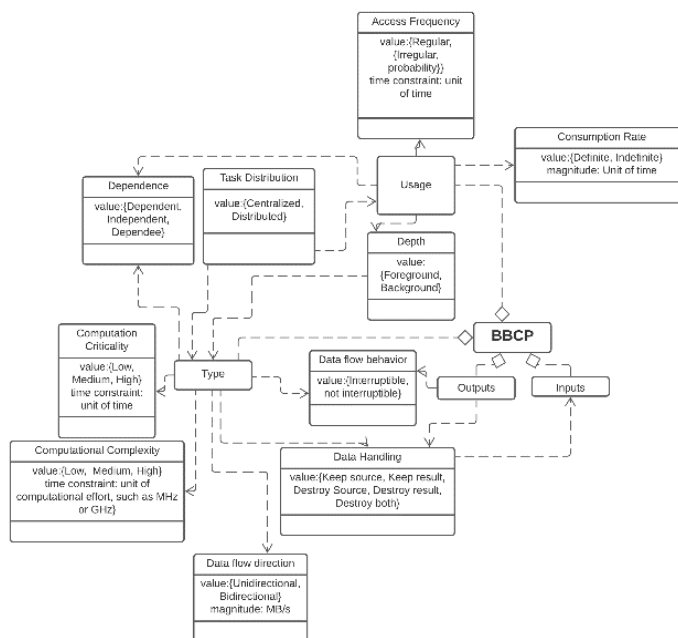
*MOTS-CLES : Profil de service, économie d'énergie, conception logicielle*

*KEYWORDS : Service Profiling, Energy consumption, software design*

---

Même si les architectes de matériel ont réussi à réduire progressivement la consommation d'énergie dans les appareils des technologies de l'information et de la communication, l'exécution des logiciels a toujours pour conséquence une consommation d'énergie. Cette situation a incité les chercheurs à élaborer un nombre des méthodologies visant à promouvoir le développement de logiciels écologiques avec de nouvelles méthodes d'évaluation dont la philosophie est de calculer les coûts énergétiques du développement et de l'exécution des. Malgré cela, elles ont été reconnues et adoptées avec un succès limité, car elles tentent d'aborder des variables hautement volatiles (comme le comportement humain) et des environnements avec des plateformes matérielles/logicielles spécifiques et des

solutions centrées sur le langage. Cela a créé un conflit entre la théorie et la pratique alors qu'autrement, une approche générique et adaptative pourrait gérer la discordance. Dans cet article, nous présentons une brève revue des recherches sélectionnées disponibles en relation avec la définition et le profilage des exigences des services pour la gestion de l'énergie, ainsi que les limites et les avantages des propositions existantes en relation avec le développement de logiciels verts. En outre, nous présentons nos progrès vers une série de propriétés permettant de définir les exigences des services et leur comportement en matière de consommation de ressources. Notre objectif final est de créer une approche appropriée pour la gestion de l'énergie à partir des phases d'analyse et de conception du cycle de vie du développement logiciel. Ces propriétés sont modélisées dans la figure ci-jointe :



Exemple pour **Access Frequency** : fréquence d'accès/de demande du service.

**Valeurs possibles** : *Regular* : prévisibilité élevée ou une fréquence spécifique au cours d'une période de temps pendant laquelle le service est invoqué/accédé/demandé ; *Irregular* : il n'y a pas d'intervalle prévisible.

**Unités de grandeur** : *Regular* : valeur sur unité de temps, par ex. : 20 accès par seconde. *Irregular* : une valeur de probabilité entre 0 et 1 concernant un intervalle de temps spécifique.



---

## Vers une approche efficace de gestion d'évolution des données graphes

**Landy Andriamampianina<sup>1,2</sup>, Franck Ravat<sup>1</sup>, Jiefu Song<sup>1,2</sup>,  
Nathalie Vallès-Parlangeau<sup>1</sup>**

1. IRIT-CNRS (UMR 5505), Université Toulouse 1 Capitole  
2 Rue du Doyen-Gabriel-Marty 31042 Toulouse, France  
*prenom.nom@irit.fr*

2. Activus Group,  
1 Chemin du Pigeonnier de la Cépière 31100 Toulouse, France  
*prenom.nom@activus-group.fr*

---

*Cet article est un résumé de l'article : Landy Andriamampianina, Franck Ravat, Jiefu Song, Nathalie Vallès-Parlangeau, Towards an Efficient Approach to Manage Graph Data Evolution: Conceptual Modelling and Experimental Assessments. RCIS 2021: 471-488.*

*MOTS-CLES : Graphe temporel, Snapshots, Evolution temporelle, Bases de données graphes.*

*KEYWORDS: Temporal graph, Snapshots, Temporal evolution, Graph data stores.*

---

Les données graphes modélisent naturellement les relations complexes entre des entités dans de nombreux domaines. Les changements au cours du temps de ces données graphes fournissent des informations contextuelles à un décideur, et ainsi nécessitent d'être intégrées aux données graphes de base. L'évolution temporelle des données graphes peut être classée en trois types, à savoir (i) *l'évolution de la topologie* (l'ajout et la suppression de nouvelles entités et relations), (ii) *l'évolution des attributs* (l'ajout et la suppression des attributs décrivant une entité ou une relation) et (iii) *l'évolution de la valeur des attributs* (changement de valeur des attributs décrivant une entité ou une relation) (Zaki et al., 2016). Nous présentons dans cet article une proposition de gestion de tous les types d'évolution des données graphes.

Les solutions existantes, nommées *graphes temporels*, ne capturent que partiellement l'évolution dans les graphes en se limitant à deux types d'évolution : la topologie et la valeur des attributs (Zaki et al., 2016). De plus, l'approche classique pour historiser les changements est la séquence de snapshots. Cette dernière ne capture pas réellement les changements dans le graphe et crée de la redondance des données qui ne changent pas (Moffitt et Stoyanovich, 2017). Afin de répondre à ces limites, notre objectif est de proposer une nouvelle approche de gestion de graphes temporels qui gère non seulement l'évolution de la topologie et de la valeur des

attributs (types (i) et (iii) mentionnés ci-dessus) mais également l'ajout et la suppression des attributs descriptifs (type (ii)). De plus, il est nécessaire de définir des règles d'implémentation qui garantissent la faisabilité et l'efficacité de la solution proposée.

La solution que nous proposons intègre un modèle conceptuel de graphe temporel composé d'*entités* et de *relations temporelles*. Une entité (ou relation) temporelle est composée d'un ensemble d'états. Chaque état contient une liste d'attributs ainsi que leurs valeurs présentes pendant une période de validité. A travers ces concepts, nous modélisons, d'une part, l'évolution de la topologie en comparant deux ensembles d'états d'entités (ou de relations) correspondant à des périodes de validité différentes, et d'autre part, l'évolution des attributs et de leurs valeurs en comparant deux états d'une même entité (ou d'une relation). Gérer les changements au niveau de chaque entité et relation permet d'éviter la redondance des données des snapshots. Au niveau logique, notre solution inclut des règles d'implémentation de notre modèle conceptuel dans un graphe de propriétés.

Notre proposition a été évaluée à travers une série d'expérimentations afin d'illustrer sa faisabilité (implémentable) et son efficacité (éviter la redondance des données et produire des temps d'exécution raisonnables). Pour ce faire, nous avons utilisé un jeu de données d'un benchmark de référence contenant les trois types d'évolution précédents. Nous l'avons implémenté à l'aide du système de gestion de bases de données graphe Neo4j. Nous avons créé 28 requêtes de référence avec une couverture complète des différents types d'analyse possibles.

Notre étude de faisabilité a permis de montrer que notre modèle temporel conceptuel est implantable sans difficulté sous Neo4j. Dans notre étude d'efficacité, nous avons implémenté notre jeu de données dans Neo4j selon trois approches de modélisation : l'approche classique des snapshots, une approche optimisée des snapshots (réduisant partiellement la redondance) et notre modèle. Nous avons comparé l'efficacité de notre proposition aux snapshots en mesurant l'utilisation du disque et le temps d'exécution des requêtes. Notre proposition permet de diviser par 12 l'utilisation du disque des implémentations basées sur les snapshots. De plus, sur la base des 28 requêtes, notre proposition permet de réduire le temps d'exécution des requêtes entre 58% et 99% suivant le type de requêtes.

En résumé, nous proposons une approche efficace de gestion de tous les types d'évolution des données graphes. La prochaine étape de nos travaux se concentre sur la manipulation des graphes temporels. L'objectif de ces travaux est d'enrichir les analyses décisionnelles via l'ajout de la temporalité dans les données graphes.

## Bibliographie

- Moffitt V. Z., Stoyanovich J. (2017). Towards sequenced semantics for evolving graphs. EDBT.
- Zaki A., Attia M., Hegazy D., Amin S. (2016). Comprehensive Survey on Dynamic Graph Models. International Journal of Advanced Computer Science and Applications. vol. 7, n°2, p. 573--582.

---

# Towards a Graph-Oriented Perspective for Querying Music Scores

Philippe Rigaux<sup>1</sup>, Virginie Thion<sup>2</sup>

1. Conservatoire National des Arts et Métiers, CEDRIC Laboratory, France

Philippe.Rigaux@cnam.fr

2. Univ. Rennes, CNRS, IRISA, Lannion, France

Virginie.Thion@irisa.fr

---

*ABSTRACT.* Sheet music scores have been the traditional way to preserve and disseminate Western classical music works for centuries. Nowadays, their content can be encoded in digital formats that yield a very detailed representation of music content expressed in the language of music notation. These encoded (digital) scores constitute an invaluable asset for digital library services such as search, analysis, clustering, and recommendations. In this paper, we propose a model of the musical content of digital score as graph data, which can be stored in a graph database management system. We then discuss the querying of such data through graph pattern queries. We also describe a proof-of-concept of the approach that allows uploading music scores in a Neo4j database, and expressing searches and analyses through graph pattern queries with the query language Cypher.

*RÉSUMÉ.* Depuis plusieurs siècles, la diffusion de la musique occidentale est assurée par la représentation des œuvres sous forme de partitions musicales. Ces partitions peuvent aujourd'hui être encodées dans des formats numériques offrant une représentation fine de leur contenu, ouvrant ainsi la voie à de nouvelles fonctionnalités. La notation musicale, même numérisée, reste cependant extrêmement complexe, notamment en raison de l'imbrication des aspects relatifs au contenu et de ceux qui décrivent la mise en forme de ce contenu. Dans cet article, nous proposons une modélisation formelle, sous forme de données graphe, du contenu purement musical des partitions numérisées. Cette modélisation vise à s'abstraire des aléas liés aux choix d'encodage et à la surcharge consécutive aux informations de mise en forme. Nous discutons ensuite des fonctionnalités d'interrogation offertes par cette représentation. Nous fournissons également une réalisation concrète de ce cadre formel, sous la forme d'une implémentation dans le système Neo4j permettant l'interrogation des données via le langage de requête Cypher.

*KEYWORDS:* Music scores, Graph databases, Data model, Pattern queries

*MOTS-CLÉS:* Partitions musicales, Base de données graphe, Modèle de données, Interrogation à base de patrons

---

## 1. Introduction

Music is an essential part of the world’s cultural heritage. Even though audio files constitute the main access channel to music works nowadays, music has been preserved and disseminated as *sheet scores* for centuries. For a part of music production, sheet scores have been – and continue to be – the most complete and accurate way to encode the composer’s intents, and to faithfully convey these intents to performers.

A sheet score is a complex semiotic object. In a single and compact layout, it combines a symbolic encoding of the music that must be produced with a sophisticated visual representation aiming at accurately representing the music content. As an illustration, Fig. 1 is the excerpt of the human-readable visualisation of a music score, extracted from a MEI dataset available in the NEUMA platform (Rigaux *et al.*, 2012; Neuma, 2022), which contains four musical voices.

Figure 1. Excerpt of *La Française*, François Couperin (NEUMA platform)

Nowadays, digital score libraries store collections of music scores, most often in the form of images i.e., scans of sheet scores. We can cite IMSLP (IMSLP, 2022) or Gallica (Gallica, 2022). In general, we expect a digital library to be more than a simple repository of digital documents, encoded in a music-agnostic format that obscures their content (Foscarin *et al.*, 2021). Services that leverage digital representation are required, and at the very least a search engine that allows retrieval of documents that match patterns of interest *by content*, the extraction of relevant patterns and features, transformation (e.g., transposition), analysis (e.g., frequent patterns, quality defaults), *etc.* To supply such intelligent services, we need a digital representation that truly encodes the *music notation* embedded in a music score, and thereby gives fine-grained access to all its components. We will call *encoded scores* such documents. The most salient formats that exist at the moment are `**kern` (KernScores, 2022; kern, 2022), MusicXML (Good, 2001; MusicXML, 2022) and MEI (Rolland, 2002; MEI, 2022).

On the over hand, the need to handle *complex* data has led to the emergence of new types of data models. In the last few years, *graph databases* (Angles, Gutierrez, 2008; Angles, 2012; Webber *et al.*, 2013; Angles *et al.*, 2017; Bonifati *et al.*, 2018) have started to attract a lot of attention in the database community. Their basic purpose is to manage data natively modelled as a graph like e.g., social networks, biological, topological databases or bibliographic databases. As the content of an encoded score is composed of highly complex information, with connected items (a note follows

another one, it belongs to a voice, it also belongs to a measure, etc.), its graph-based representation seems a relevant approach to leverage existing encodings to a true data model apt at supporting analytic operations, including searches.

In this paper, we propose two complementary contributions. First, we model the content of encoded scores as a graph, which can then be stored in a graph database management system, and processed with graph pattern queries. We expose the formal structure of the model, illustrate the querying mechanism, and discuss its expressiveness. Second, we propose a proof-of-concept of the approach based on a tool called MUSYPHER, which allows producing the graph-based representation of an encoded score and then uploading it in a Neo4j graph database, in order to query such data with the concrete query language Cypher.

The paper is organized as follows. Section 2 presents a review of literature focused on recently proposed music score content models. Section 3 introduces our graph-based data model. Section 4 studies the properties of graph patterns queries. Section 5 then presents our implementation. Section 6 concludes and draws some perspectives of this work.

## 2. Related work

When modelling an encoded score, the literature provides two main approaches. The first one consists in a tree-shaped model. The second one sees the music content as a collection of time series, modelling musical events performed over a time period. In the following, we succinctly present these two approaches.

**Tree-based modelling.** Driven by the need of interoperability between systems and tools, semi-structured models have emerged for representing (Western) music scores. Widespread ones are MusicXML (Good, 2001; MusicXML, 2022) and MEI (Rolland, 2002; MEI, 2022). Their common characteristics is to encode the complete content of sheet scores (including graphical specifications that dictate how the notation content has to be visually rendered), and to organize, in the form of a n-ary ordered tree, the music events (e.g., notes, lyrics) according to the measure they belong to. Generic database query languages may be applied to such documents like XPath or XQuery (Ganseman *et al.*, 2008; Fournier-S’niehotta *et al.*, 2018). However, their complexity, mix of several concerns and the fact that a same information can be encoded with many syntactic variants hinder the definition of robust music-oriented query languages. We take as a starting point in the present paper the recent proposal of (Zhu *et al.*, 2022) to normalize the hierarchical structure of music score.

**Times-series modelling.** Another perspective consists in seeing the music content as time series of musical events (called “voices”). The *ScoreAlg* algebra, defined in (Fournier-S’niehotta *et al.*, 2018), is based on such a perspective. A voice is a function whose domain is a time measurement (relevant division of time for timestamping the musical events e.g., a time unit that represents the smallest interval between two musical events) and co-domain is the set of musical events. Voices are then synchro-

nized in order to form a complex score. *ScoreAlg* provides a closed form language based on operators for manipulating music scores information. Such a language, powerful but based on complex abstractions, may be difficult to use for the average user who is familiar with music notation but not the digitized encoding of music scores, e.g., a music performer (who retrieves music scores in order to play music with his band), a music analyst (who searches for similar patterns in the parts of a music score), or a musicologist (who conducts a philological study on Rameau’s compositions including all their variants over time).

**Graph-based modelling of the rhythm.** Driven by the need to enhance an Optical Music Recognition (ORM) process, (Jin, Raphael, 2015) proposes an approach that represents a music content from a rhythmic perspective. The rhythm itself is a graph where each node models the occurrence of a rhythmically relevant symbol (note, rest, and bar line). Each symbol has a duration (a dotted eighth has duration of  $3/16$ , while a bar line has a duration of 0). Nodes are connected either by their order of occurrence inside a voice (a symbol follows another one) or by coincidence edges (symbols of different voices are connected if they share the same onset time). This model focuses on the alignment of the music sheet symbols in order to refine music symbols recognition.

Our graph-based model follows the trend, explored in the above-mentioned papers, to abstract the content of encoded scores in order to expose a robust representation, focused on music content, and cleared from side information related to representation purposes. The graph representation captures both a tree-based representation that exposes the hierarchical nature of a score, and a time series perspective, close to the intuitive understanding of a music score as a flow of sound events.

### 3. Graph-based modelling of encoded scores

We now turn to the definition of a graph-based data model for modelling score content. It relies heavily on principles taken from music notation, seen as an expressive formal language that provides a powerful basis for modelling music content. Our model gives an abstract vision of digital music documents as structured objects, focusing on music content, cleared from side information related to representation purposes, and supports query functionalities developed in the forthcoming sections. It will be illustrated with the German anthem, *Das Lied der Deutschen*, composed by *Joseph Haydn* in 1797 (Haydn, 1797). The notation of this example is shown on Fig. 2.



Figure 2. First notes of the German anthem, *Das Lied der Deutschen* by *Joseph Haydn* (1797)

To state it in a nutshell, we model music information as a mapping from a structured temporal domain to a set of (musical) *facts*. The temporal domain is a hierarchical structure, called *rhythmic tree*, that partitions a finite time range in non-overlapping

intervals. Each interval defined by a leaf of the rhythmic tree is associated with a music fact. Together they constitute a *musical event* (i.e., a fact that occurs during a specific temporal interval).

### 3.1. Preliminaries: the property graph model

In a graph database management system, the schema is a graph (nodes are entities and edges are relations between entities), and data is handled through graph-oriented operations and type constructors (Angles, Gutierrez, 2008; Angles, 2012; Webber *et al.*, 2013; Angles *et al.*, 2017). Our modelling relies the *property graph* data model (Angles *et al.*, 2017; Bonifati *et al.*, 2018), where nodes and edges may embed data in *properties* (key-value pairs). In terms of vocabulary, we assume the existence of the pairwise disjoint sets: a set  $\mathcal{V}$  of *nodes*, a set  $\mathcal{E}$  of *edges*. We also consider a set *Lab* of labels, and a set *Prop* of *properties* (a.k.a. *property keys*), and a set *Val* of *values*.

**DEFINITION 1 (Property graph).** — A *property graph*  $\mathcal{G}$  is a tuple  $(V, E, \rho, \lambda, \sigma)$  where (1)  $V \in \mathcal{V}$  is a finite set of nodes; (2)  $E \in \mathcal{E}$  is a finite set of edges; (3)  $\rho : E \rightarrow (V \times V)$  is a total function assigning to each edge an ordered pair of nodes (where  $\rho(e) = (n_1, n_2)$  indicates that  $e$  is an edge going from  $n_1$  to  $n_2$ ); (4)  $\lambda_V : V \rightarrow \mathcal{P}(\text{Lab})$  is a partial function assigning a set of labels to the nodes of  $V$ ; and  $\lambda_E : E \rightarrow \text{Lab}$  is a total function assigning a label to each edge of  $E$ ; (Without ambiguity,  $\lambda$  refers either to  $\lambda_V$  or to  $\lambda_E$  according to the domain of the assigned element.) (5)  $\sigma : (V \cup E) \times \text{Prop} \rightarrow \text{Val}$  is a partial function assigning property-value pairs to nodes of  $V$  and edges of  $E$  (where  $\sigma(n, p) = v$  (resp.  $\sigma(e, p) = v$ ) indicates that the node  $n$  (resp. edge  $e$ ) has a property  $p$  with the value  $v$ ).

Classical notions come with the definition of a property graph  $\mathcal{G} = (V, E, \rho, \lambda, \sigma)$ .

**DEFINITION 2 (Path).** — A *path*  $p$  in  $\mathcal{G}$  is a sequence  $n_1 l_1 n_2 l_2 n_3 \dots n_{k-1} l_{k-1} n_k$  where  $\{n_1, \dots, n_k\} \subseteq V$  and  $\{l_1, \dots, l_{k-1}\} \subseteq \text{Lab}$  and  $k \geq 1$  and each  $(n_i, l_i, n_{i+1})$  s.t.  $i \in \{1, \dots, k-1\}$  is an edge of  $\mathcal{G}$ .<sup>1</sup> Path  $p$  is said to connect  $n_1$  to  $n_k$ , and its size is  $|p| = k - 1$ . Any non empty path that is a subsequence of  $p$ , is a *subpath* of  $p$ . Path  $p$  is a *cycle* iff  $k \geq 2$  and  $n_1 = n_k$ . Path  $p$  is *cyclic* iff one of its subpaths is a cycle, otherwise  $p$  is said to be *acyclic*.

**DEFINITION 3 (Paths).** — Let  $n_1$  and  $n_2$  be two nodes of  $V$ . We denote by  $\text{Paths}(n_1, n_2)$  the set of acyclic paths that connect  $n_1$  to  $n_2$ .

### 3.2. The domain of musical facts

Several fact domains can be envisaged to describe a musical object. Due to space limitation, we will focus here on the main one, *sounds*. A sound can be characterized

1.  $(n_i, l_i, n_{i+1})$  is an edge of  $\mathcal{G}$  means that there is  $e \in E$  such that  $\rho(e) = (n_i, n_{i+1})$  and  $\lambda(e) = l_i$ .

by many properties, including intensity, timbre and frequency. In the language of music notation, a finite set of frequencies, or *pitches*, is used to refer to the sounds usable in a musical piece. We follow the designation of the International Standards Organization (ISO) for enumerating the pitches. In this designation, each pitch is referred to by a pitch class  $P$  (a letter A, B, C, D, E, F, or G), an index  $I$  in  $[1, 7]$ , and an optional accidental  $a$  in  $\{\sharp, \flat\}$  (*sharp* and *flat* accidentals). We will thus model a sound as a triple  $\{class : P, octave : I, accidental : a\}$ , and represent it as symbol of the form  $P[a]I$ . Addition of other relevant properties (e.g., intensity) is trivial.

Graphically (i.e., in music scores), frequency levels are materialized by groups of horizontal lines (called staves) and pitches are represented by black or white heads vertically positioned on staves. The first pitch in the score of Fig. 2 is a  $C5$ , followed by a  $D5$ , an  $E5$ , etc. Music is also made of silences (or *rests*), and we thus add the *rest symbol*  $r$  to the domain. The German anthem starts with a rest, graphically represented by a small rectangle.

Finally, in conventional music notation, sounds can be “tied” (graphically represented as curves over the heads, such as in the first measure of Fig. 2). We add the *continuation symbol*  $_$  to our domain to represent ties. We obtain the domain of musical facts.

**DEFINITION 4** (Domain of (atomic) musical facts). — *The domain  $\mathcal{F}_{\mathcal{M}}$  of musical facts consists of:*

1. *the set of sound facts  $P[a]I$ ,  $P \in \{A, B, C, D, E, F, G\}$ ,  $a \in \{\sharp, \flat\}$ ,  $I \in [1, 7]$ ,*
2. *the rest fact, noted  $r$ ,*
3. *the continuation fact, noted  $_$ .*

Facts are represented as nodes with properties. Fig. 3.(a) is a fact node that models an instance of an  $A4\sharp$  note. The node has three properties, namely `class` (this property has the value A for the node), `octave` (having the value 4) and `accid` (having the value `sharp`, which models a  $\sharp$  alteration). In the following, such a node will be depicted by a compact representation embedding the core property values in the node itself (Fig. 3.(b)).



Figure 3. Fact node: instance of an  $A4\sharp$  note

We can derive some important notions from musical facts. An *interval* is a distance between two sounds, physically characterized by the ratio of their respective frequencies. A ratio of 1 denotes a *unison*, a ratio of 2 an *octave*. The octave is the fundamental interval that structures symbolic music representation. In Western music notation, an octave range is divided in 12 *semi-tones*. This defines a scale, called *chromatic*, with 12 *steps*, corresponding each to exactly one semi-tone. A *chromatic interval* is the number of semi-tones between two pitches.



### 3.3. Temporal organization of facts

A music piece is a temporal organization of sounds inside a bounded time range. Musical facts cannot be assigned to any timestamp but fall on a set of positions that defines a discrete partitioning of this range. More precisely, this partition results from a recursive decomposition of temporal intervals, yielding a rhythmic organization which is inherently hierarchical.

In Western music notation, a music piece is divided in *measures* (graphically represented as vertical bars on Fig. 2), and a measure contains one or more *beats*. Beats can in turn be divided into equal units (i.e., sub-beats). Further recursive divisions often occur, generating a hierarchy of pulses called *metrical structure*. The *time signature*, a rational number (in our example 4/4 denoted by **C**), determines the preferred decomposition. A 4/4 measure consists of 4 beats, and each beat is one quarter (graphically, a black note ♩) long. Still in the context of a 4/4 time signature, the preferred decomposition of a measure, is into 4 sub-intervals (some other partitions are possible, although less likely), beats are preferably partitioned in two quavers (graphically, a ♪), themselves (generally) partitioned in semi-quavers (♩), etc. For other meters (e.g., 3/4, 6/8), temporal decomposition follows different patterns. However, in all cases, the metrical structure can be represented as a *rhythmic tree* based on the following principles: i) the time range of the piece is divided in equal-sized measures, ii) each measure is recursively divided according to metric rules, and iii) each leaf of the tree corresponds to a musical fact that occurs during the sub-interval defined by the position of the leaf.

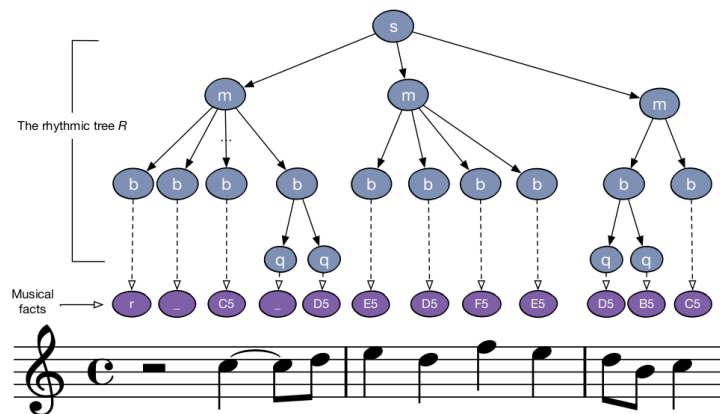


Figure 4. Modelling the German anthem, with its rhythmic tree and associated facts

Fig. 4 illustrates our structure: a rhythmic tree that temporally organizes the musical facts. The part with blue nodes represents the rhythmic tree. The root (*s*) corresponds to the whole music piece. The level under the root is made of *measure nodes*. Each measure is itself decomposed in beats, quavers, etc., according to the required temporal decomposition. The first measure for instance is decomposed in four beats, the latest being itself decomposed in two quavers.

Each leaf of the rhythmic tree is linked to a musical fact (the purple nodes). The first beat of the first measure is a *rest* fact, the second is a *continuation* fact (extending the previous rest), the third is a *sound* fact (a  $C5$ ). The fourth beat is divided in two facts: a continuation of the  $C5$ , and a  $D5$  fact. This structure summarizes our modelling, and the basis for building a graph representing music content. Note that we have only been discussing so far of *monodic* music i.e., a single flow of sounds. Respecting a well-established vocabulary in the field of music notation, we will call such a structure a *voice*.

**DEFINITION 5 (Voice).** — *A voice is a pair  $(R, \mathcal{M})$ , where  $R$  is a rhythmic tree and  $\mathcal{M}$  is mapping from the leaves of  $R$  to the set of musical facts  $\mathcal{F}_{\mathcal{M}}$ .*

Given a voice, we can easily infer several properties that will serve as a basis for the querying process. Let us start with the durations and temporal positions of the facts, determined from the rhythmic tree thanks to the following definition.

**DEFINITION 6 (Temporal partition).** — *Let  $I = [\alpha, \beta[$  be a time range and  $R$  a rhythmic tree. The temporal partitioning of  $I$  with respect to  $R$  assigns an interval  $itv_I(N)$  to each node  $N$  of  $R$  as follows.*

1. *If  $N$  is the root of  $R$ ,  $itv_I(N) = I$*
2. *If  $N$  is of the form  $N(N_1, \dots, N_n)$ ,  $itv_I(N) = [\alpha_N, \beta_N[$  is partitioned in  $n$  sub-intervals of equal size  $s = \frac{\beta_N - \alpha_N}{n}$  each:  $itv_I(N_i) = [\alpha_N + (i-1) \times s, \alpha_N + i \times s[$*

Given a time range  $I$ , the leaf level of a rhythmic tree  $R$  defines a partitioning of  $I$  as a set of non-overlapping temporal intervals. Now, consider a voice  $(R, \mathcal{M})$ . Each leaf  $l$  of  $R$  covers an interval  $itv_R(l)$ , whose duration is naturally that of its associated fact  $\mathcal{M}(l)$ . Together,  $itv_R(l)$  and  $\mathcal{M}(l)$  constitute a *musical event*.

**DEFINITION 7 ((Musical) event).** — *Let  $I$  be a time range, and  $V = (R, \mathcal{M})$  a voice. If  $l$  is a leaf of  $R$ , then the pair  $(itv_I(l), \mathcal{M}(l))$  is a musical event.*

We adopt the following convention to represent temporal values: the duration of a measure is 1, it extends over interval  $[0, 1[$ , and the music piece range is  $I = [0, n[$ ,  $n$  being the number of measures. Both the duration and interval of a node result from the recursive division represented by the tree. If the meter is  $4/4$ , the duration of a half note for instance is  $\frac{1}{2}$ , the duration of a beat is  $\frac{1}{4}$ , etc. Turning back to Fig. 4, the first event is  $(I_1, r)$ , with  $I_1 = [0, \frac{1}{4}[$  (the first beat of the first measure). The second event is  $(I_2, -)$ , with  $I_2 = [\frac{1}{4}, \frac{2}{4}[$  (we recall that it represents a continuation of the rest for one beat). The third event is  $([\frac{2}{4}, \frac{3}{4}[, C5)$ , the fourth is  $([\frac{3}{4}, \frac{7}{8}[, -)$ , etc.

### 3.4. Polyphonic Music

The representation of polyphonic music simply consists of a set of voices sharing a same number of measures. Fig. 5 gives an illustration (the same theme, with a bass part added). In terms of modelling, we have two choices. Either we simply model a polyphonic piece as a set of voices  $\{V_1, V_2, \dots, V_n\}$ , with the constraint that all the upper level (measures) of their respective rhythmic trees are similar. Or we “factorize”



Figure 5. German anthem, with two voices

these upper levels, representing the common sequence of measures, and we manage individual subtrees for each measure and each voice.

### 3.5. Graph based representation of voices

We now turn to the graph-based modelling of a voice. The graph is almost directly obtained from the voice structure thanks to the following transformations:

1. An edge is added between adjacent measures: this allows to navigate from one measure to the other.
  2. Intermediate levels between the level of the measures and the leaf levels are removed: measures are directly connected to leaves.
  3. Since the previous simplification removes the structural information that determines the temporal features (intervals and durations) of the leaves (see Def. 7), we add these features explicitly in the graph: an edge is added between two adjacent leaves, with a duration property.
  4. The content of a voice may be reached either from the root of the rhythmic tree (the rhythmic tree perspective) or from its first event (the time series perspective).
- Any relevant information that qualifies a voice as a whole may be attached to the voice node e.g., the instrument or the name of the voice.

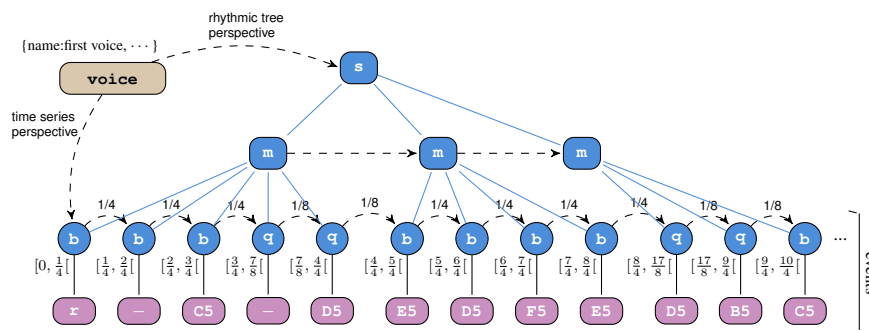


Figure 6. Graph-based representation of a voice

One obtains, from the voice of Fig. 4, the graph of Fig. 6. It can be seen as a compacted representation of the voice structure, with pre-computed important values (e.g., durations), and additional edges that help navigating the graph.

#### 4. Querying a graph-based music score content

Let us now consider the problem of querying the graph-based modelling of an encoded score. We first recall the basic notion of property graph pattern. A *graph pattern* is classically defined as a graph where variables and conditions can occur (Barceló, 2013; Angles *et al.*, 2017). Intuitively, it defines a shape that has to be found in data.

**DEFINITION 8** (Graph pattern syntax). — *Let  $Var_{nodes}$  and  $Var_{lab}$  be distinct sets of node variables and label variables respectively. A graph pattern query is a tuple of the form  $(V, E, \rho, \lambda, \sigma, Cond_n, Cond_e)$  where variables can occur on nodes and on edge labels<sup>2</sup>, and  $Cond_n$  (resp.  $Cond_e$ ) denotes Boolean conditions over property values of the elements of  $V$  (resp.  $E$ ) (e.g., if  $f_1$  is a node variable, then a condition could be  $f_1.class = E \wedge f_1.octave = 4$ , which indicates in our context that the node mapping in  $f_1$  must be a E4 fact note).*

For the sake of querying a music content, we extend the notion of graph pattern query to the relevant musical vocabulary, which allows using, in the pattern conditions, the notions of interval between two notes and distance between two events.

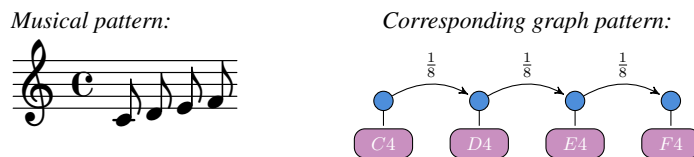


Figure 7. Sequence of C4-D4-E4-F4

In its simplest form, a graph pattern denotes a path. Fig. 7 and 8 are some examples of such derived patterns. In these patterns, nodes and edges are free variables, and conditions may be attached to the pattern. The pattern of Fig. 7 aims at retrieving the sequences of eighth notes C4-D4-E4-F4 that appear in data. The pattern contains conditions over the pitch class and octave of the notes, and over the duration of the events.

The pattern of Fig. 8 aims at retrieving the occurrences of two notes following a C4 one. Fact1 and Fact2 are node variables. A condition is attached with Fact1 that must be a E note. Another condition is attached to the pattern concerning the adequacy of the octaves of Fact1 and Fact2. There is no other condition over the notes (for instance, there is no condition over the duration of each of the three notes).

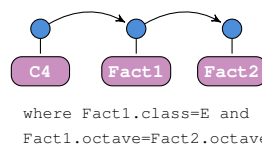


Figure 8. Notes following a C4

2. This means that  $(V, E, \rho, \lambda, \sigma)$  is a property graph such that  $V \in \mathcal{V} \cup Var_{nodes}$  and  $\lambda : (V \cup E) \rightarrow Lab \cup Var_{lab}$ .

A query may also have the form of a more complex graph pattern. For instance, the complex pattern  $P_{twoG4}$  of Fig. 9.(a) aims at retrieving two  $G4$  notes that belong to the same measure in the same voice (but the notes are not necessary adjacent). The wavy relationship between the events denotes an arbitrary path. Another complex pattern is  $P_{poly}$  (Fig. 9.(b)), which aims at retrieving the occurrence of two notes ( $n3$  and  $n4$  on the figure) played during the time of another one ( $n1$  on the figure) in another voice.

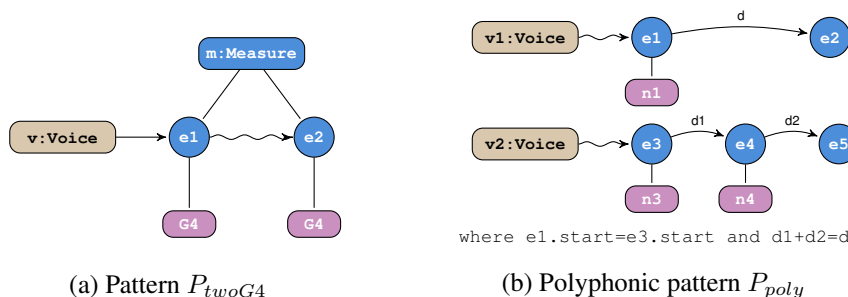


Figure 9. More complex graph patterns

Formally, the interpretation (evaluation)  $\llbracket P \rrbracket_{\mathcal{G}}$  of a graph pattern query  $P$  over a graph database  $\mathcal{G}$  consists in finding all the subgraphs of  $\mathcal{G}$  that are homomorphic to  $P$  (Barceló, 2013).

DEFINITION 9 (Graph pattern interpretation). — *The answer  $\llbracket P \rrbracket_{\mathcal{G}}$  of a pattern query  $P$  over a graph  $\mathcal{G}$  is the set of subgraphs  $\{g \in \mathcal{P}(\mathcal{G}) \mid g \text{ "matches" } P\}$ . A subgraph  $g$  "matches"  $P$  iff there is a homomorphism  $h$  from the nodes and labels variables of  $P$  to  $\mathcal{G}$  and each node (resp. label)  $h(v)$  satisfies its associated conditions  $Cond_n$  (resp.  $Cond_e$ ).*



Figure 10. J.S. Bach cantate BWV111, 6th movement

In order to illustrate the interpretation of a graph pattern query, we consider the polyphonic music score of Fig. 10, denoted by BWV111 in the following. The encoding of this music score is available (MEI format) in the J.S. Bach collection of the NEUMA platform (Neuma, 2022). The music score contains four voices: *Soprano*, *Alto*, *Tenor* and *Bass*. Five measures appear on Fig. 10, numbered from 0 to 4.

Fig. 11 presents the graph-based representation of BWV111, restricted to the beginning of the *Soprano* and *Alto* voices. This graph respects the model proposed in

Section 3. The rhythmic tree, common to all the voices, appears in blue. Fact nodes appear in purple. The *Soprano* and *Alto* voice nodes appear on the left. With each voice is associated the time series of events that compose the voice content. The edges of the time series embed, in property values, the *duration* of each event, expressed here in the time unit of the *beat*. The events embed their time range in additional properties named *start* and *end* (by lack of space, these values do not appear in the figure but can easily be inferred from the duration value carried by the edges of the time series). This graph is denoted by  $\mathcal{G}_{\text{BWV111}}$  in the following.

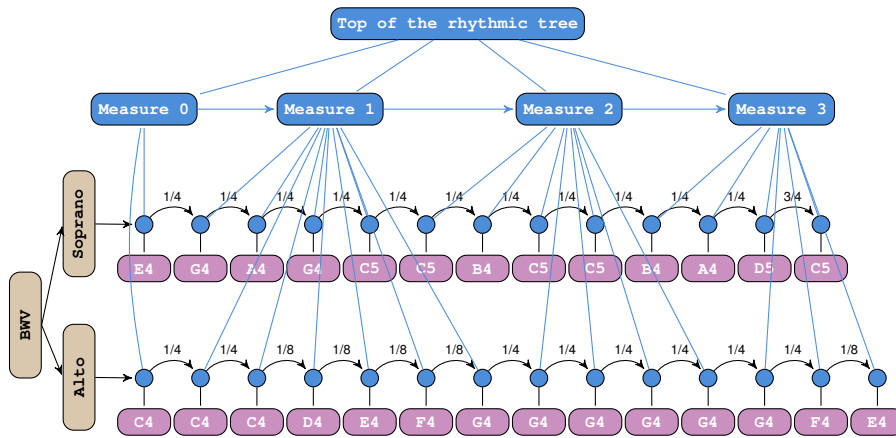


Figure 11. ( $\mathcal{G}_{\text{BWV111}}$ ) Graph-based modelling of BWV111

Let us also consider the graph pattern  $P_{twoG4}$  of Fig. 9.(a). This pattern matches into several subgraphs of  $\mathcal{G}_{\text{BWV111}}$ , including a subgraph that contains the first and third notes of the measure 1 for the voice *Soprano*, and another subgraph that contains the first and second notes of the measure 2 for the voice *Alto*. These two corresponding subgraphs of  $\mathcal{G}_{\text{BWV111}}$  belong to  $\llbracket P_{twoG4} \rrbracket_{\mathcal{G}_{\text{BWV111}}}$ . Note that other answers that match  $P_{twoG4}$  can be found in  $\mathcal{G}_{\text{BWV111}}$ .

**Expressiveness, limitations and cost.** In terms of expressiveness of the model, the graph-based model itself -from a static point of view- implements the hierarchical modelling of the rhythm and the time series oriented perspective of (Fournier-S'niehotta *et al.*, 2018). It also captures the core structure of a tree-based representation of a music content (XML-based formats). Some rhythmic features like those considered in the model of (Jin, Raphael, 2015) are also captured, limited to musical events (those that "produce a sound") and their coincidence (series of such events in a voice).

The problem of querying a graph has been studied in the literature (Barceló, 2013; Angles *et al.*, 2017). In terms of query expressiveness, as a tree is special case of a graph, the navigational queries (e.g., XPath-like) that applied over a tree-based rep-

resentation can be applied to the graph-based structure (Libkin *et al.*, 2013).<sup>3</sup> Graph patterns queries, whose evaluation is based on graph pattern matching, are a very expressive formalism, not surprisingly more expensive in terms of evaluation cost. Such a cost depends on the form of the query pattern (Wood, 2012; Barceló, 2013; Barceló *et al.*, 2014; Angles *et al.*, 2017), going from  $O(|\mathcal{G}| \cdot |P|)$ , where  $|\mathcal{G}|$  is the size of the data graph and  $|P|$  is the size of the pattern, for a pattern having the "simple" form of a regular path query (a path connecting two nodes where the edge is labelled by a regular expression), to an NP-complete problem for the most general case of a complex pattern that may contain a cycle.

## 5. Implementation

We implemented a proof of concept of the proposed framework, in the Neo4j graph database management system (Neo4j, 2022), for querying a music score through the Cypher (Cypher, 2022) query language. The implementation is based on a software tool called MUSYPHER, developed for our project.<sup>4</sup> MUSYPHER makes possible to process a MEI (XML-based) file in order to translate its music content into a Neo4j graph database that respects the graph-based representation proposed in Section 3.

In order to illustrate the implementation, we consider again the music score of Fig. 10 (p. 11), whose digitized content was initially available in the MEI format in the J.S. Bach collection of the NEUMA platform (Neuma, 2022). This music score was translated in its graph-based representation (see Fig. 11) and loaded in a Neo4J database.

We now consider the querying of such data "in practice". Let us return to the graph pattern  $P_{twoG4}$  of Fig. 9.(a). In the Cypher language, a query based on  $P_{twoG4}$  could be expressed by Query  $Q_{twoG4}$  given in Fig. 12. The shape of the graph pattern is declared in the `MATCH` clause (lines 1 to 6). In the Cypher formalism, a graph pattern is defined in the manner of the ASCII art, where the graphic symbol `( )` denotes a node, which may contain information of the form `variable:Type`, and the symbol `-[expr]->` defines the form of a connection (`expr` is either an edge label, or a regular expression denoting a path). Expressions of the form `{attr1:value1, attr2:value2, ...}` are additional conditions over properties. The `RETURN` clause (lines 7 to 9) contains the elements of interest that should be rendered as a result. Here, we retrieve for each answer subgraph: the number of the measure, the name of the voice, and the initial MEI identifier of the elements<sup>5</sup> that was considered as a relevant property in our context, to be transferred into the graph.

3. Of course, the literature proposes a flavour of navigational query languages.

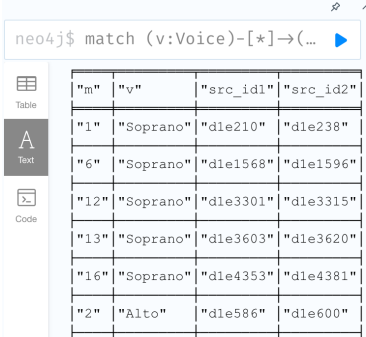
4. MUSYPHER is a Java application, based on a DOM parsing of the initial XML encoded score. The current version of the tool is available at <https://www-shaman.irisa.fr/musypher>

5. Knowing the MEI identifiers, we can highlight the elements in a rendering tool like Verovio (Pugin *et al.*, 2014).

```

1 MATCH
2   (v:Voice)-[*]->(e1:Event)-[*]->(e2:Event),
3   (m:Measure)--(e1),
4   (m:Measure)--(e2), // the same measure m
5   (e1)--(note1{class:'g',octave:4}),
6   (e2)--(note2{class:'g',octave:4})
7 RETURN
8   m.number AS m, v.name AS v,
9   e1.id AS src_id1, e2.id AS src_id2

```



The screenshot shows a Neo4j query interface. The query is: `neo4j$ match (v:Voice)-[*]->(...`. The results are displayed in a table with the following columns: `m`, `v`, `src_id1`, and `src_id2`. The table contains 8 rows of data.

"m"	"v"	"src_id1"	"src_id2"
"1"	"Soprano"	"die210"	"die238"
"6"	"Soprano"	"die1568"	"die1596"
"12"	"Soprano"	"die3301"	"die3315"
"13"	"Soprano"	"die3603"	"die3620"
"16"	"Soprano"	"die4353"	"die4381"
"2"	"Alto"	"die586"	"die600"

Figure 12. Query  $Q_{twoG4}$  and the result of its evaluation by Neo4j over  $\mathcal{G}_{BWV111}$

A part of the result of  $Q_{twoG4}$ , evaluated by the Neo4j query evaluator engine over  $\mathcal{G}_{BWV111}$ , is given in the Neo4j desktop screenshot placed at the right of Fig 12. Among the answers of the table, the first and sixth answers correspond to the answer subgraphs previously discussed according to the interpretation of  $P_{twoG4}$  over  $\mathcal{G}_{BWV111}$  (see section 4, p. 12).

## 6. Conclusion and perspectives

In this paper, we proposed a graph-based data model for modelling the musical content of digital scores. This model captures the two perspectives of a music score content: (i) as a metrical structure that starts from a rhythmic organization of the temporal range, and associates a fact with each leaf of the rhythmic tree, and (ii) as a time series of events, which is close to the intuitive understanding of a music score as a flow of sounds. We discussed the querying of such data through graph pattern queries. We also presented a proof-of-concept of the approach that allows uploading music scores in a Neo4j database (using a tool called MUSYPHER), and expressing searches and analyses through graph pattern queries with the query language Cypher.

This work opens many perspectives. Some of them concern the extension of the model. First, we considered only the domain of sound facts, but other domain can be added to enrich the description of music information, following the same principle: such information is a mapping from a rhythmic organization of time to facts. The domain of *syllables* for instance allows to model lyrics in vocal music; one could add domains of semantic facts (timbre, texture, intensity) to model in general *annotations* that qualify temporal fragments of a music piece. A second kind of extension could address more sophisticated musical facts, such as harmonic sounds, ornaments, performance directives, etc.

Concerning the querying process, we plan to conduct a more detailed theoretical study of the cost of querying, as advanced results of the literature concerning the cost of graph pattern queries exhibited several classes of tractable queries (e.g., patterns



that do not contain a cycle (Barceló, 2013)). These classes should be mapped into musical patterns.

Finally, this representation could be combined with previous results for managing data quality in graph databases (Pivert *et al.*, 2020; Rigaux, Thion, 2017), in order to detect data quality problems in the music score content (Foscarin *et al.*, 2021). And finally, in terms of implementation, the MUSYPHER tool is still under development. A current short-term development consists in integrating MUSYPHER in the NEUMA platform, in the form of a module allowing the graph-based querying of the available collections.

#### *Acknowledgements*

*The authors thank Clément Van Straaten for his work on the MUSYPHER tool.*

#### **References**

- Angles R. (2012). A comparison of current graph database models. In *Proc. of the intl. conf. on data engineering (icde) workshops*, p. 171-177.
- Angles R., Arenas M., Barceló P., Hogan A., Reutter J. L., Vrgoc D. (2017). Foundations of modern query languages for graph databases. *ACM Comput. Surv.*, Vol. 50, No. 5, pp. 68:1–68:40.
- Angles R., Gutierrez C. (2008). Survey of graph database models. *ACM Computing Surveys (CSUR)*, Vol. 40, No. 1, pp. 1–39.
- Barceló P. (2013). Querying graph databases. In *Proc. of the ACM Symposium on Principles of Database Systems (PODS)*, pp. 175–188.
- Barceló P., Libkin L., Reutter J. L. (2014). Querying regular graph patterns. *J. ACM*, Vol. 61, No. 1, pp. 8:1–8:54.
- BnF. (2022). *Bibliothèque numérique de la BnF*. <https://gallica.bnf.fr/>. (Accessed Feb. 2022)
- Bonifati A., Fletcher G. H. L., Voigt H., Yakovets N. (2018). *Querying graphs*. Morgan & Claypool Publishers.
- Foscarin F., Rigaux P., Thion V. (2021). Data Quality Assessment in Digital Score Libraries. The GioQoso Project. *Intl. Journal on Digital Libraries*, Vol. 22, No. 2, pp. 159-173.
- Fournier-S'niehotta R., Rigaux P., Travers N. (2018). Modeling Music as Synchronized Time Series: Application to Music Score Collections. *Information Systems*, Vol. 73, pp. 35–49.
- Ganseman J., Scheunders P., D'haes W. (2008). Using XQuery on MusicXML databases for musicological analysis. In *Proc. of the intl. conf. on music information retrieval, (ismir)*, pp. 433–438.
- Good M. (2001). The virtual score: Representation, retrieval, restoration. In, p. 113-124. W. B. Hewlett and E. Selfridge-Field, MIT Press.
- Haydn J. (1797). *Das lied der deutschen*. (Lyrics by August Heinrich Hoffmann von Fallersleben)

- Humdrum. (2022). *Representing Music Using \*\*kern*. <https://www.humdrum.org/guide/ch02/>. (Accessed Feb. 2022)
- IMSLP. (2022). *Intl. Music Score Library Project*. <https://imslp.org>. (Accessed Feb. 2022)
- Jin R., Raphael C. (2015). Graph-based rhythm interpretation. In *Proc. of the intl. society for music information retrieval conf. (ismir)*, pp. 343–349.
- KernScores. (2022). <http://kern.ccarh.org/>. (Accessed Feb. 2022)
- Libkin L., Martens W., Vrgoč D. (2013). Querying Graph Databases with XPath. In *Proceedings of the Intl. Conf. on Database Theory ICDT*, pp. 129–140.
- Music Encoding Initiative. (2022). <http://www.music-encoding.org>. (Accessed Feb. 2022)
- MusicXML. (2022). <http://www.musicxml.org>. (Accessed Feb. 2022)
- Neo Technology. (2022). *The Neo4j Manual*. <https://neo4j.com/developer>. (Accessed Feb. 2022)
- Neo4j web site. (2022). [www.neo4j.org](http://www.neo4j.org). (Accessed Feb. 2022)
- NEUMA. (2022). <http://neuma.huma-num.fr>. (Accessed Feb. 2022)
- Pivert O., Scholly E., Smits G., Thion V. (2020). Fuzzy quality-aware queries to graph databases. *Information Sciences*, Vol. 521, pp. 160–173.
- Pugin L., Zitellini R., Roland P. (2014). Verovio: A library for Engraving MEI Music Notation into SVG. In *Proc. of the Intl. Society for Music Information Retrieval (ISMIR)*, pp. 107–112.
- Rigaux P., Abrouk L., Audéon H., Cullot N., Davy-Rigaux C., Faget Z. *et al.* (2012). The design and implementation of neuma, a collaborative digital scores library - requirements, architecture, and models. *Intl. Journal on Digital Libraries*, Vol. 12, No. 2-3, pp. 73–88.
- Rigaux P., Thion V. (2017). Quality Awareness over Graph Pattern Queries. In *Proc. of the Intl. Database Eng. & Applications Symp. (IDEAS)*.
- Rolland P. (2002). The Music Encoding Initiative (MEI). In *Proc. of the Intl. Conf. on Musical Applications Using XML*, p. 55-59.
- Webber J., Robinson I., Eifrem E. (2013). *Graph databases*. O'Reilly Media.
- Wood P. T. (2012). Query languages for graph databases. *SIGMOD Rec.*, Vol. 41, No. 1, pp. 50-60.
- Zhu T., Fournier-S'niehotta R., Rigaux P., Travers N. (2022). A Framework for Content-based Search in Large Music Collections. *Big Data and Cognitive Computing*, Vol. 23, No. 6.

---

## Détection de signaux faibles : une méthode basée sur les graphlets

Hiba Abou Jamra, Marinette Savonnet, Éric Leclercq

Laboratoire d'Informatique de Bourgogne – EA 7534  
Université de Bourgogne Franche-Comté – 9 Avenue Alain Savary, F-21078  
Dijon - France  
Hiba\_Abou-Jamra@etu.u-bourgogne.fr

---

**RÉSUMÉ.** Cet article est un résumé de l'article (Abou Jamra et al., 2021) portant sur la détection des signaux faibles afin d'aider les experts métier dans leur prise de décision. La détection des signaux faibles permet aux décideurs ou à la population de se préparer de manière appropriée à des événements futurs. Nous nous intéressons particulièrement aux données issues des réseaux sociaux qui peuvent être vues comme un graphe. Nous cherchons une signature des signaux faibles à l'aide d'une étude topologique en utilisant comme outil opératoire les graphlets. Grâce à leurs formes et leurs tailles prédéfinies, les graphlets présentent aussi l'avantage de pouvoir être interprétés par les experts métier.

**ABSTRACT.** Weak signals detection allows decision makers or population to prepare appropriately for upcoming events. We are particularly interested in data from social networks that can be seen as a graph. We are looking for a signature of weak signals by means of a topological study using graphlets as an operational tool. Thanks to their predefined shapes and sizes, graphlets can be interpreted by domain experts.

---

**MOTS-CLÉS :** SIGNAUX FAIBLES, TOPOLOGIE DES RÉSEAUX, GRAPHLETS

**KEYWORDS:** WEAK SIGNALS, NETWORK TOPOLOGY, GRAPHLETS

---

Dans les organisations actuelles qui gèrent des environnements complexes, être capable d'anticiper les discontinuités et les événements futurs permet de répondre à une menace ou de saisir une opportunité. Le volume et la diversité des informations produites empêchent les acteurs responsables de voir les signaux qui peuvent avertir d'événements importants à venir et inconnus *a priori*. Par conséquent, prévoir ces signaux futurs et agir correctement à temps est un défi difficile à relever. Cependant, la notion du signal faible n'est pas définie avec précision dans la littérature, car les auteurs utilisent des termes différents pour la désigner, notamment : "signe d'avenir", "alerte précoce", "indicateur de changement", "wild cards".

Nous adoptons la première définition des signaux faibles proposée par Ansoff en 1975 qui les définit comme les premiers symptômes de discontinuités stratégiques

agissant comme une information d’alerte précoce, de faible intensité, pouvant être annonciatrice d’une tendance ou d’un évènement important. Les signaux faibles possèdent plusieurs propriétés quantifiables qui permettent leur caractérisation et aident à leur détection. De ces caractéristiques nous retenons : **fragmentaire, visibilité faible, peu ou pas familier, utilité faible et fiabilité faible.**

Dans un monde complexe où l’information circule rapidement, les entreprises doivent surveiller leur environnement et ne plus être centrés mais au contraire avoir une vision périphérique. L’information produite par les réseaux sociaux est une bonne source pour trouver des signaux faibles annonciateurs de tendance ou de menace pour l’entreprise. C’est pour cela que nous nous intéressons à la détection des signaux faibles cachés dans les discours issus des réseaux sociaux et plus particulièrement de Twitter.

La plupart des approches de détection des signaux faibles étudient l’émergence de mots-clés à l’aide des techniques de *text-mining*. Nous proposons une autre voie en analysant la topologie du réseau, afin de trouver une propriété quantifiable qui peut être caractéristique du signal faible. C’est pourquoi nous avons choisi les graphlets (Pržulj et al., 2004) comme description opératoire pour détecter les signaux faibles. En effet, les graphlets répondent aux caractéristiques des signaux faibles : ce sont de petits patterns (fragments d’un graphe), qui pris seuls sont peu visibles et de faible utilité apparente. Nous étudions l’évolution des graphlets au cours du temps en utilisant leurs vitesse et accélération comme marqueurs de diffusion et d’amplification du signal. De plus, la taille des graphlets et leurs formes prédéfinies ainsi que les orbites, c’est-à-dire les positions ou les rôles des nœuds dans les graphlets, facilitent leur interprétation par des experts métier. Ces caractéristiques nous permettent de surveiller et d’expliquer le rôle d’individus influents détectés à travers des algorithmes de centralité comme le Page Rank.

Les expériences que nous avons réalisées ont comme objectifs l’étude de cas réels, la reproductibilité, et la vérification que des faux positifs ne sont pas détectés quand l’évènement est déjà prévu et ponctuel. Les résultats de nos expérimentations ont conforté notre hypothèse que les graphlets peuvent être considérés comme signature d’un signal faible. Ils permettent à la fois d’automatiser la tâche de détection des signaux faibles dans un gros volume de données tout en laissant une place à l’interprétation par des experts gommant ainsi l’effet « boîte noire » que pourrait avoir une méthode entièrement automatisée. Les perspectives consistent à : 1) mener de nouvelles expériences sur d’autres types et sur de plus grands réseaux ; 2) voir si les graphlets peuvent déterminer des moments critiques (transitions de phase) dans des systèmes dynamiques complexes comme la finance.

### **Bibliographie**

- Abou Jamra H., Savonnet M., Leclercq É. (2021). Detection of event precursors in social networks: A graphlet-based method. In International conference on Research Challenges in Information Science, p. 205–220.
- Pržulj N., Corneil D. G., Jurisica I. (2004). Modeling interactome: scale-free or geometric Bioinformatics, vol. 20, no 18, p. 3508–3515.

---

# Spygraph : un robot d'exploration léger dédié à l'analyse de graphes d'hyperliens

Robert Viseur<sup>1</sup>

1. FWEГ - Université de Mons  
Service de Technologies de l'Information et de la Communication  
17, place Warocqué, B-7000 Mons  
[robert.viseur@umons.ac.be](mailto:robert.viseur@umons.ac.be)

---

*RÉSUMÉ.* Spygraph est un programme écrit en Python permettant l'exploration d'un ensemble de sites web de manière à en extraire les hyperliens. Facilement installable, configurable et exécutable, il outille l'exploration d'un écosystème local au travers des sites web des acteurs qui le composent. Basé sur une structure de données simplifiée en SQLite, il permet de configurer l'exportation des hyperliens vers différents outils utiles à l'analyse tels que LibreOffice.org Calc (tableur) ou Gephi (analyse de graphe). Dès lors, il facilite la découverte de nouveaux acteurs au travers des domaines liés et de l'analyse de leurs relations en ligne. Spygraph se positionne donc comme une solution légère, adaptée aux graphes de petite taille, en complémentarité avec des solutions plus complexes comme Hyphe.

*ABSTRACT.* Spygraph is a program written in Python allowing the exploration of a set of websites in order to extract hyperlinks. Easily installable, configurable and executable, it enables the exploration of a local ecosystem through the websites of the organisations that make it up. Based on a simplified data structure in SQLite, it allows to configure the export of hyperlinks to different tools useful for the analysis such as LibreOffice.org Calc (spreadsheet) or Gephi (graph analysis). It therefore facilitates the discovery of new actors through linked domains and the analysis of their online relationships. Spygraph is thus positioned as a light solution, adapted to small graphs, in complementarity with more complex solutions like Hyphe.

*MOTS-CLÉS :* analyse de graphe, robot d'exploration, Gephi.

*KEYWORDS:* graph analysis, crawler, Gephi.

---

## 1. Contexte et présentation générale

Spygraph a été développé dans le cadre du projet Interreg [FabricAr3v](#)<sup>1</sup>. L'objectif était de disposer d'un outil léger, basé sur des composants libres et *open source*, permettant l'exploration d'un écosystème local, transfrontalier, composé d'acteurs économiques actifs dans la fabrication numérique, au travers des hyperliens reliant leurs sites web, et en complément d'approches quantitatives (p. ex. sondages) ou qualitatives (p. ex. entretiens). À cette fin, des métriques d'analyse de graphe ont été utilisées avec le logiciel dédié [Gephi](#). (cf. Hansen *et al.*, 2010 pour

---

<sup>1</sup> Cf. <https://fabricar3v.eu/>.

une présentation des méthodologies d'analyse de graphe et Figure 1 pour un exemple de graphe tiré du projet [FabricAr3v](#)).

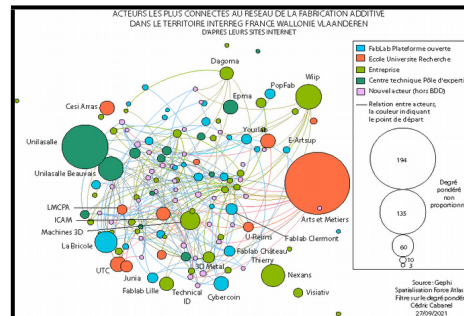


Figure 1. Résultat de l'analyse de graphe

Spygraph est donc un programme dédié à l'exploration d'un ensemble de sites web (robot) et à la production d'une base de données d'hyperliens. L'exportation dans un format de fichier adapté permet d'ensuite réaliser une analyse des relations entre sites web à l'aide d'un logiciel libre commun comme [LibreOffice.org Calc](#) (tableur) ou spécialisé comme [Gephi](#)<sup>2</sup> (analyse de graphe). Spygraph est développé en [Python](#) et s'appuie sur une base de données [SQLite](#). Il utilise des bibliothèques courantes telles que [urllib](#) (manipulation et lecture des URLs) et [Beautiful Soup](#)<sup>3</sup> (lecture des documents HTML) de manière à en faciliter l'installation, l'extension et l'usage. L'outil se veut léger et flexible ; il n'ambitionne pas de remplacer des solutions plus complexes, adaptées aux plus grands graphes, comme [Hyphe](#)<sup>4</sup>.

## 2. Utilisation du programme

Le script principal, actuellement testé sous Ubuntu Linux, doit être lancé en ligne de commande et accepte des paramètres d'entrée. L'avancée du processus d'exploration peut être suivie dans un terminal de type Bash sous Ubuntu Linux. Un cycle d'exploration complet peut être lancé à l'aide d'un script Bash :

```
#!/bin/bash
python3 run.py init <monprojet>
python3 run.py crawl inside <monprojet>
python3 run.py export csv <monprojet>
python3 run.py export graph <monprojet>
```

Le programme exporte un fichier DOT ainsi qu'un fichier CSV. Ce dernier peut être importé dans un classeur LibreOffice.org Calc dès lors que le nombre de lignes n'excède pas les limites du format [Open Document Format](#), soit 1.048.576 lignes dans le cas des fichiers portant l'extension « ods »<sup>5</sup>.

La configuration du programme se fait, d'une part, lors de l'exécution de la ligne de commande (cf. Tableau 1), d'autre part, au sein d'un fichier de configuration. La configuration se fait au sein d'un fichier textuel (`run.cfg`), compatible avec le

<sup>2</sup> Cf. <https://gephi.org/>.

<sup>3</sup> Cf. <https://beautiful-soup-4.readthedocs.io/>.

<sup>4</sup> Cf. <https://github.com/medialab/hyphe>.

<sup>5</sup> Cf. <https://wiki.documentfoundation.org/Faq/Calc/022>.

module Python [Config Parser](#), permettant de paramétrer la profondeur de l'exploration, le nombre maximal d'URLs traitées avant arrêt, le *timeout*, ainsi que les requêtes SQL utilisées par défaut pour l'exportation des données.

Tableau 1. Paramètres de ligne de commande

init	Création de la base de données et initialisation de la table « urls_init » avec la liste d'URLs d'amorçage.	
crawl	Exécution de l'exploration sur base de la liste d'URLs.	
	inside	Exploration des URLs associées aux domaines fournis dans la liste d'amorçage uniquement.
	outside	Exploration de toutes les URLs trouvées (mode divergent).
export	Exportation des données issues de la base de données une fois l'exploration terminée.	
	csv	Exportation des liens entre pages au format CSV.
	graph	Exportation des liens entre pages au format DOT.

Sur un ordinateur [Dell E6540](#), équipé d'un disque dur SSD, d'un processeur Intel Core i7-4810MQ (2,80GHz) et de 16GB de RAM, tournant sous Ubuntu Linux, connecté à un réseau universitaire (fournisseur d'accès [Belnet](#)), la vitesse d'exploration, en fonctionnement *monothread*, est de l'ordre de 75 URLs par minute. Sur une table d'URLs comportant plus d'un million d'entrées, l'exécution d'une requête de sélection impliquant des jointures et/ou des regroupements n'excède pas 5 à 10 secondes.

### 3. Structure de base de données

La base de données est composée de trois tables (cf. Tableau 2). La première contient la liste des URLs d'amorçage. Chaque URL est associée au *fully qualified domain name* (FQDN) et au nom de domaine calculés (p. ex. *commons.wikimedia.org* et *wikimedia.org*). La seconde contient la liste des URLs découvertes. La troisième contient la liste des liens entre pages, identifiés par les pages sources et destinations, associées à leurs domaines respectifs. *Stricto sensu* la table « urls » contient des URIs incluant les URLs mais aussi des [schémas d'URIs](#) (comme « *mailto* », « *tel* » ou « *javascript* »). Chaque entrée y est associée à un champ « *iscrawled* », indiquant si l'URI a été visitée, un champ « *isignored* », indiquant si l'URI a été ignorée (cas des schémas autres que les schémas d'URL), et un champ « *iserror* », indiquant si l'ouverture ou la lecture de la page ont déclenché une erreur. Ces champs permettent l'identification d'éventuels problèmes lors de l'exploration. Ils ont par exemple permis d'identifier, sur les premières versions du robot, des problèmes d'instabilité réseau et de vérification de certificats SSL.

Tableau 2. Structure de la base de données

urls_init	urls	links
url fulldomain domainname	url fulldomain domainname iscrawled iserror isignored	hyperlink_from hyperlink_to fulldomain_from fulldomain_to domainname_from domainname_to

L'utilisation de la base de données [SQLite](#) permet une navigation aisée dans les résultats de l'exploration que l'on travaille sous Windows, macOS ou GNU/Linux. L'analyse des résultats de l'exploration, la sélection des données ainsi que le filtrage des données lors de l'exportation peuvent en effet être réalisés en interrogeant directement la base de données en utilisant la console SQL du logiciel libre [DB Brower for SQLite](#)<sup>6</sup>. À l'aide du SQL, il est ainsi possible de rapidement découvrir de nouveaux sites et de dresser des statistiques sur base des résultats de l'exploration : nombre d'URLs utilisées en amorçage, nombre d'URIs traitées, nombre d'URLs explorées, nombre de liens extraits, nombre de liens extraits sans doublon, nombre d'URIs identifiées... Les novices en SQL peuvent se rabattre sur un outil de tableur, alimenté par le fichier CSV exporté, tel que Microsoft Excel ou LibreOffice.org Calc (p. ex. calcul du nombre d'URL par domaine unique et détection des domaines découverts lors de l'exploration).

#### 4. Perspectives

Premièrement, la lecture des documents pour en extraire les hyperliens pourrait être complétée par une extraction de terminologie. Cela permettrait, d'une part, la génération automatique d'un nuage de mots-clefs pour un ensemble de domaines sélectionnés par requêtes (cf. [Viseur, 2014a](#) pour un retour d'expérience avec des outils NLP *open source*), d'autre part, le filtrage des URLs, lors de l'exportation, sur base d'une liste de mots-clefs. Deuxièmement, le contenu des pages pourrait être utilisé pour créer un index plein texte à l'aide d'une technologie *open source* telle que [Lucene](#)<sup>7</sup> ou un de ses portages (p. ex. [PyLucene](#)) (cf. [Viseur, 2012](#), et [Viseur, 2014b](#), pour des exemples). Troisièmement, et plutôt que de privilégier la couverture maximale d'un ensemble de sites web prédéfinis, l'objectif étant de découvrir les acteurs en interaction sur un territoire donné au travers des liens entre leurs sites web respectifs, le robot pourrait intégrer des techniques de « *focused crawling* » afin de réduire la taille du corpus d'URLs (cf. [Micarelli & Gasparetti, 2007](#) pour une introduction) sans altérer les qualités exploratoires du robot.

#### Bibliographie

- Hansen D., Shneiderman B. Smith M.A. (2010). *Analyzing social media networks with NodeXL: Insights from a connected world*. Morgan Kaufmann.
- Micarelli A. & Gasparetti F. (2007). Adaptive focused crawling. In *The adaptive web* (pp. 231-262). Springer, Berlin, Heidelberg.
- Viseur R. (2014a). Automating the Shaping of Metadata Extracted from a Company Website with Open Source Tools. *International Journal of Advanced Computer Science and Applications*.
- Viseur R. (2014b). Initial Results from the Study of the Open Source Sector in Belgium. *International Symposium on Open Collaboration*, Berlin, Germany.
- Viseur R. (2012). Create a Specialized Search Engine: The Case of an RSS Search Engine. *International Conference on Data Technologies and Applications*, Rome, Italie.

<sup>6</sup> Cf. <https://sqlitebrowser.org/>.

<sup>7</sup> Cf. <https://lucene.apache.org/>.



---

## Une méthode pour la modélisation de la variabilité des indicateurs de performance des processus intégrée dans des modèles de processus variables

Diego DIAZ<sup>1,2</sup>, Mario CORTES-CORNAX<sup>1</sup>, Agnès FRONT<sup>1</sup>, Cyril LABBE<sup>1</sup>, David FAURE<sup>2</sup>

1. Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France  
*prenom.nom@univ-grenoble-alpes.fr*

2. Groupe INCOM&COSI+, 2 Av. de Vignate, 38610 Gières  
*Diego.Diaz@incom-sa.fr, David.Faure@incom-sa.fr*

---

REFERENCE DE L'ARTICLE INTERNATIONAL. Cet article est un résumé de l'article :  
Diego Diaz, Mario Cortes Cornax, Agnès Front, Cyril Labbé, David Faure:  
*A Method for Modeling Process Performance Indicators Variability Integrated to Customizable Processes Models. RCIS 2021: 72-87.*

MOTS-CLES : Indicateur de performance, Famille de processus, Variabilité

KEYWORDS: Process Performance Indicators, Process Families, Variability

---

Très souvent, les organisations doivent adapter leurs processus métier, et en conséquence les indicateurs de performance de leurs processus (Process Performance Indicators - PPI), selon les besoins des clients ou les nouvelles réglementations. La définition et le calcul des indicateurs de performance dans le contexte de processus métier fortement variables sont d'autant plus consommateurs en temps, sources d'erreurs et dépendants de la connaissance et du savoir-faire d'experts métiers. Du point de vue de leur conception, les processus et les indicateurs de performance sont généralement modélisés indépendamment, de façon souvent découplée : les modèles qui supportent la variabilité et l'adaptation des processus métiers tels que les familles de processus métiers (La Rosa et al., 2017) ne supportent pas la variabilité des indicateurs de performance ; inversement, les approches permettant de modéliser la variabilité des indicateurs de performance telles que PPINOT (del-Rio-Ortega et al., 2019) abordent la variabilité des indicateurs dans le contexte de modèles de processus prédéfinis et ne sont pas intégrés dans des modèles de processus variables.

La méthode PPIC (Process Performance Indicator Calculation) permet de modéliser la variabilité des indicateurs de performance dans des processus métiers variables. Elle étend la méthode BPFM (Business Process Feature Model) (Cognini et al., 2016). Nos contributions sont à trois niveaux : I) la méthode PPIC composée de 5 étapes et basée sur un modèle de caractéristiques pour les indicateurs de performance PPICT (Process Performance Indicator Calculation Tree) (Diaz, 2020) ; II) un métamodèle et sa syntaxe concrète, permettant de modéliser la variabilité des indicateurs de performance ; III) un prototype support à la méthode, développé sur la plateforme ADOxx.

La méthode PPIC est composée de 5 étapes proposées en Figure 1 : (1) construction du modèle de caractéristiques PPICT, (2) conception des indicateurs de performance, (3) intégration du PPICT dans le modèle BPFM, (4) configuration du modèle de caractéristiques, (5) vérification de la conformité entre les configurations des indicateurs de performance et du modèle de processus. L'application de la méthode sur un cas d'étude réel dans le domaine de la gestion des contrats publics d'eau et d'électricité est présentée ainsi que les résultats d'une évaluation qualitative centrée utilisateurs.

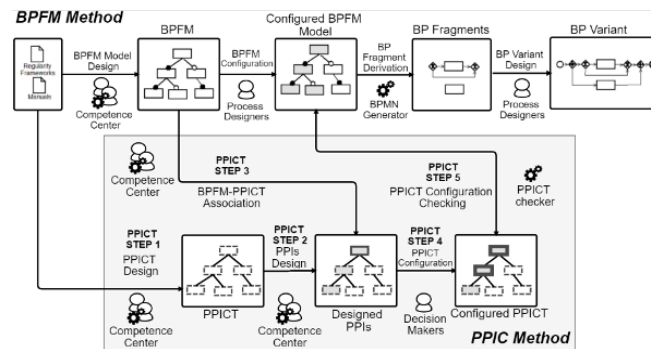


Figure 1. La méthode PPIC, extension de la méthode BPFM

## Bibliographie

- Cognini, R., Corradini, F., Polini, A., Re, B. (2016) Business process feature model: an approach to deal with variability of business processes. In: Domain-Specific Conceptual Modeling. pp. 171–194. Springer.
- del-Río-Ortega, A., Resinas, M., Durán, A., Bernárdez, B., Ruiz-Cortés, A., Toro, M (2019) Visual PPINOT: A graphical notation for process performance indicators. Business & Information Systems Engineering. 61, 137–161.
- Diaz, D. (2020) Integrating PPI Variability in the Context of Customizable Processes by Extending the Business Process Feature Model. In: 2020 IEEE 24th International Enterprise Distributed Object Computing Workshop (EDOCW). pp. 80–85. IEEE..
- La Rosa, Marcello, Van Der Aalst, W.M.P., Dumas, M., Milani, F.P (2017) Business process variability modeling: A survey. ACM Computing Surveys (CSUR). 50, 2.

---

# Traitement des événements complexes pour une gestion proactive des instances d'un processus métier

**Abir Ismaïli-Alaoui<sup>1,2</sup>, Khalid Benali<sup>1</sup>, Karim Baïna<sup>2</sup>**

1. *Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France*

2. *Rabat IT Center, ENSIAS, Université Mohammed V, Rabat, Maroc*

*abir.ismaili-alaoui@univ-lorraine.fr , khalid.benali@loria.fr, karim.baina@um5.ac.ma*

---

**RÉSUMÉ.** *Le management des processus métier (Business Process Management – BPM) est vu comme la solution adéquate qui aide les organisations à s'adapter aux évolutions stratégiques, organisationnelles et techniques. Le BPM aide à avoir plus de visibilité et de contrôle sur les processus métier, de la modélisation des processus métier jusqu'à l'exécution et l'optimisation en cas de besoin. Cependant, la transformation digitale et l'essor de plusieurs nouvelles technologies telles que les big data, le cloud computing, et l'internet des objets (IoT), etc, engendrent de nouveaux défis, dans le domaine du BPM, liés entre autres aux ressources (humaines / machines) limitées et à la nécessité d'optimiser davantage l'utilisation de ces ressources, ainsi que l'exploitation des données et événements. La notion d'événement est très importante dans le BPM. En effet, au cours de son exécution, un processus métier génère beaucoup de données (event data/event logs). En plus, ses instances sont généralement déclenchées par des événements. C'est pourquoi, l'analyse et le traitement de ces événements ont une grande importance dans l'amélioration de la gestion de processus métier. Dans ce papier, nous essayons de montrer comment les techniques et méthodes proposées par le domaine du traitement des événements complexes (Complex Event Processing – CEP) peuvent être exploitées afin d'améliorer certains aspects du BPM, principalement pour les processus métier basés sur les événements.*

**ABSTRACT.** *Business Process Management – BPM helps organizations adapt to strategic, organizational and technical changes, as it enables more visibility and control of business processes and their activities, from modeling to execution and optimization if required. However, the new digitalized era and the rise of several new technologies such as big data, fast data, cloud computing, Internet of Things (IoT), etc, implies new business process challenges linked basically to limited (human / machine) resources and the need to use these resources in an optimal way. During process execution a lot of data and event data are generated. Moreover, business process instances are mostly triggered by events. That is why analysing and processing these events is of great importance in improving business process management. In this paper, we try to explain that techniques and methods proposed by the field of Complex Event Processing (CEP) can be exploited to improve certain aspects of BPM, mainly for event-driven business processes. In fact, these techniques can help to*

*process and filter these events, to optimize their management during the execution of business process instances. A case study is presented and the obtained results from our experimentations demonstrate the benefit of our approach and allowed us to confirm the efficiency of our assumptions.*

*Mots-clés : Gestion des Processus Métier, traitement des événements complexes, événements, Internet des objets, proactivité, priorité.*

*KEYWORDS: Business Process Management, Complex Event Processing, Event, IoT, Proactivity, priority*

---

## **1. Introduction**

L'approche processus s'est imposée de plus en plus aux entreprises à partir des années 80, donnant ainsi un nouveau modèle d'organisation et un nouveau mode de fonctionnement. Face à un environnement changeant et concurrentiel, les approches traditionnelles qui traitent l'entreprise comme un milieu clos ne sont plus adaptées. En effet, l'approche processus est une approche systémique qui vise à transformer la structure hiérarchique et verticale d'une organisation en une structure transversale qui a comme objectif ultime la satisfaction des clients internes et externes à l'entreprise, surtout dans cette ère de la transformation digitale (Baiyere et al., 2020, Lederer et al., 2017). C'est une méthode d'analyse et de modélisation destinée à assurer un travail collectif, afin de maîtriser et d'améliorer l'efficacité et le bon fonctionnement d'un organisme. Cette méthode se base principalement sur la notion des processus, d'ailleurs le niveau de performance d'une entreprise est lié directement à la performance des processus et à la qualité de leurs modèles. En effet, une bonne gestion des processus métier dans une entreprise peut avoir un impact très positif sur l'efficacité et le bon fonctionnement de ses activités, car elle permet à l'entreprise d'avoir une vision claire sur ses objectifs, afin de mieux répondre aux exigences de compétitivités qui augmentent exponentiellement et donc atteindre un très bon degré de performance.

Le management des processus métier est vu comme un gestionnaire de workflow basé sur les activités, qui aide les organisations à surveiller de façon optimale le fonctionnement de leurs activités. Lorsqu'une entreprise met en place l'ensemble des moyens proposés par la démarche du BPM, c'est dans le but d'avoir plus de visibilité et de contrôle sur leurs processus métiers et les interactions de ces derniers, afin d'être en mesure de les modéliser, les piloter, les améliorer et les optimiser continuellement. Par conséquent, gagner en termes d'agilité, de flexibilité et de performance, c'est-à-dire rendre l'entreprise « *capable de s'adapter rapidement à un environnement d'affaires changeant tant au niveau des défis que des opportunités* » (Cummins, 2009).

Ces dernières années, les entreprises se trouvent face à une vague de nouveaux facteurs redéfinissant le marché et bousculant le BPM traditionnel. Parmi ces nouveaux facteurs on trouve la quantité de données vertigineuse, provenant, avec une très grande vélocité (big data), de différentes sources hétérogènes (des interactions internes ou externes de l'entreprise, IoT, etc). Ces données (data / event data) doivent être bien analysées et exploitées afin d'en extraire des données à forte

valeur ajoutée qui peuvent aider l'entreprise dans son processus de prise de décision. Cependant, les outils traditionnels proposés par la méthode du management des processus métiers présentent différentes limites concernant le traitement, la fouille et l'analyse des données et des événements, à un temps quasi réel. En étudiant plusieurs modèles de processus métier, nous réalisons que dans la plupart des cas, les processus métier fonctionnent de façon réactive (Mousheimish, 2017), ce qui n'est pas suffisant face à ces nouveaux changements radicaux ou incrémentaux rencontrés par les entreprises. Une anticipation précoce est cruciale pour éviter la survenue du problème ou pour y réagir rapidement et efficacement. Ce manque de proactivité et de prévisibilité est remarquable dans les trois principales étapes du cycle de vie du BPM (Van Der Aalst, 2013): l'étape de la conception, l'étape d'implémentation et l'étape d'exécution. L'état de l'art du BPM le prouve, car nous constatons l'émergence de nouveaux concepts, liés à la proactivité, qui commencent à être utilisés dans le glossaire du BPM, comme par exemple: la gestion proactive des processus métier (proactive business process management) (Mousheimish, 2017), (Ismaili-Alaoui et al., 2018b), la prévision des processus ou BPM orientée futur (process forecasting or future-oriented BPM) (Poll et al., 2018), processus métier sensibles au contexte (context-aware business processes) (Anastassiou et al., 2016), juste à titre d'exemple. Par conséquent, la transition vers un processus métier proactif et adaptable, au lieu d'un processus métier réactif, est devenu obligatoire pour chaque entreprise. La gestion des événements et des instances d'un processus métier déclenchées par ces événements illustrent bien l'importance de cette transition. En effet, en BPM traditionnel, une entreprise doit penser à l'avance à la planification et l'allocation des ressources (humaines ou machines), pour exécuter les activités de son processus, afin d'assurer un service efficace, même pendant les périodes de pointe. Cependant, avec les approches de gestion existantes, l'entreprise peut soit faire face à un sous-approvisionnement (lorsqu'il y a une sous-estimation des ressources nécessaires, les processus métier ne peuvent pas être exécutés) ou à un sur-approvisionnement (les ressources planifiées à l'avance pour couvrir les périodes de pointe ne sont pas utilisées) (Armbrust *et al.*, 2010), (Schulte *et al.*, 2015), soit ne pas bien gérer les niveaux de priorité des instances à exécuter et dans ce cas faire occuper une ressource par une instance moins prioritaire (Ismaili-Alaoui et al., 2018a; 2018b).

À cet égard, nos travaux de recherche visent à contribuer à cette émergence de la proactivité dans le BPM, en mettant l'accent sur la phase de la gestion des événements et de l'ordonnancement des instances dans le BPM. Cependant, l'ordonnancement chronologique des instances d'un processus métier et la planification préalable des ressources, empêchent la transition vers une exécution proactive, en particulier lorsque les instances de ce processus sont lancées par des événements générés en permanence par des objets connectés (IoT devices). Pour traiter un problème aussi complexe et réaliser une planification proactive des instances d'un processus basée sur la priorité, nous devons analyser l'événement qui déclenche ces instances pour estimer (voir même prédire) au préalable leur priorité. Nous cherchons donc à introduire une approche basée sur le traitement des événements complexes (CEP), qui assure un traitement et une analyse et un filtrage, en temps réel, des événements générés à partir de différentes sources, afin d'extraire

des informations utiles et de détecter des modèles d'événements complexes en temps réel. Cette approche vise à assurer une gestion proactive des événements, puis une exécution basée sur les priorités pour les instances de processus métier lancées par ces événements.

La combinaison entre CEP et BPM n'est pas récente. En fait, elle a été largement utilisée pour contrôler et surveiller les processus métier en temps réel afin d'améliorer l'efficacité des opérations métier en gardant une trace de ce qui se passe et en alertant les utilisateurs dès qu'un problème est détecté. Plusieurs approches et solutions basées sur l'intégration du CEP avec BPM ont été proposées, soit pour la phase de conception ou la phase de l'exécution d'un processus métier, (Redlich et Gilani, 2011), (Weidlich *et al.*, 2011), (Reinartz *et al.*, 2015), (Koetter et Kochanowski, 2015), (Soffer *et al.*, 2017) juste pour en nommer quelques-unes. Cependant, l'ordonnancement et la gestion des événements n'ont pas encore bénéficié de cette intégration BPM/CEP, dans le but de gérer efficacement les instances d'un processus métier déclenchées par ces événements afin d'assurer une exécution proactive.

La suite du papier est organisée comme suit : Dans la section suivante, nous présentons un exemple de motivation et notre contexte de travail. Dans la troisième section, nous présentons les travaux similaires qui ont proposé des approches pour améliorer l'ordonnancement des processus. La section 4 décrit l'approche et la méthodologie proposées. La section 5 présente le bilan des résultats obtenus. La dernière section conclut le papier et esquisse ses perspectives.

## 2. Contexte de Travail et Scénario de Motivation

Notre travail de recherche se base sur une étude de cas qui s'inscrit particulièrement dans le cadre de l'économie des seniors ou la Silver Economie. Cette dernière représente en effet, un secteur industriel récent, lancé officiellement en 2013, en France (Bernard *et al.*, 2013), dans le but de développer des services personnalisés basés sur les nouvelles technologies, afin de créer plus d'autonomie chez les personnes âgées et assurer un pas de plus *vers* des sociétés amies des aînés « age-friendly societies ».

D'après l'institut National de la Statistique et des Études Économiques (INSEE), si les tendances démographiques récentes se maintiennent, la France métropolitaine comptera 73,6 millions d'habitants au 1<sup>er</sup> janvier 2060, soit 11,8 millions de plus qu'en 2007. Le nombre de personnes de plus de 60 ans augmentera, à lui seul, de plus de 10 millions (En 2060, en France, une personne sur trois aura ainsi plus de 60 ans). Chaque année, plus de 2 millions de personnes âgées de plus de 65 ans chutent, et une personne sur 2 âgée de plus de 80 ans en est victime. Première cause de décès accidentel chez les plus de 65 ans, la chute impacte souvent la condition physique mais également la condition psychologique de la personne. Perte de confiance, peur de tomber à nouveau, repli sur soi, les conséquences d'une chute sont multiples, elles sont souvent graves et liées à l'importance de la blessure et à l'état de santé de la personne. Les chutes (qui sont à l'origine de nombreuses fractures et de milliers de décès chaque année), concernent en premier lieu les plus de 75 ans. Près d'un tiers des personnes âgées autonomes et la moitié de celles

vivant en maison de retraite ou établissement de soins de longue durée sont victimes d'au moins une chute par an. Que la blessure qui en résulte soit sérieuse ou non, l'accident n'en reste pas moins traumatisant pour la personne qui, le plus souvent, est dans l'incapacité de se relever seule. Un véritable syndrome post-chute peut alors s'installer. (d'après le site informatif et spécialiste du domaine de la Silver économie <http://www.silvereco.fr/>).

Toutes ces études et ces statistiques montrent alors que les personnes âgées présentent un risque accru de perte d'autonomie et sont de plus en plus sujet à des chutes. Il est donc primordial de créer de nouveaux outils permettant de les assister au quotidien en cas d'incident. L'intervention rapide après une chute, aidé par un détecteur de chute par exemple, pourrait éviter 26% des hospitalisations, soit 160 M€ et 9 400 décès par an. Les entreprises actives dans ce marché de la Silver économie, proposent des solutions de détection de chutes qui peuvent être classées selon trois catégories :

- **Solutions Passives** : La personne concernée doit appuyer sur un bouton d'alerte ou de signalisation en cas d'incident.

- **Solutions Actives** : La personne concernée porte un capteur / un objet connecté (accéléromètre, signaux biologiques, ...) ou des détecteurs environnementaux (présence, sol, portes...). En cas de changement ou de variation particulière des signaux, le dispositif déclenche automatiquement une alerte,

- **Solutions Actives-Vidéo** : cette solution se base principalement sur la vidéo-surveillance. Le capteur vidéo dans ce cas là analyse le comportement de la personne concernée et déclenche l'alerte si nécessaire.

Considérons, dans ce contexte, une entreprise qui offre un service innovant pour la détection et la prévention de chute chez les personnes âgées. Cette entreprise propose deux services pour ses clients basés sur deux types de solutions de détection de chutes, selon la particularité et également la préférence du client en question :

- **Service 1** (Solutions Actives Vidéo) : basé sur la vidéo-surveillance et les services de télé-assistance, qui édite un système de détection automatique de chutes pour seniors fragiles, ce système est constitué d'une chaîne d'analyse temps réel de flux de données (data streaming) relatives à des mouvements physiques de personnes âgées, en service gériatrique ou en maisons de retraites.

- **Service 2** (Combinaison de Solutions Actives et Solutions Actives Vidéo) : généralement les solutions actives qui utilisent différents types de capteurs (sols actifs, bracelets, détecteur de présence,...) sont considérées comme des solutions « aveugles », C'est-à-dire qu'elles ne permettent pas de savoir sur la base du signal d'alerte reçu si la chute est réelle ou non. De ce fait, les secours sont des fois dérangés, avec les coûts et conséquences associés, sans aucune certitude. D'où l'intérêt de combiner ce type de solution avec de la vidéo-surveillance. En effet, dans certains cas, seule l'image délivrée par les vidéo-détecteurs de chutes, permet de faire une levée de doute (écarter les fausses alertes) et par conséquent d'éviter des interventions inutiles et de minimiser le coût global de la prestation « Aide à la

détection de chute » incluant les dispositifs de détection, de l'intervention et de l'assistance de la personne. Cependant, la combinaison de ces deux solutions offre plus de précision d'une part. D'une autre part, les données collectées par les dispositifs utilisés dans les deux solutions combinées, enrichissent davantage les analyses effectuées en amont, ce qui permettent d'aller plus vers une solution proactive (une prévention de la chute) et non pas seulement réactive (une simple détection de la chute).

Les deux services proposés dans ce cas d'étude assurent une assistance tout en maintenant une autonomie et une liberté de déplacement sans entrave. Ils contribuent à l'autonomie protégée et sécurisée et à la qualité de vie des seniors fragiles. Ces services sont destinés à la fois aux seniors à domiciles ou en établissements d'hébergement pour personnes âgées dépendantes (EHPAD).

Dans cet article, nous allons nous intéresser seulement au premier type de service qui est basé sur la vidéo-surveillance. Ce service est basé sur l'analyse des images alertes envoyées en cas d'incident par les caméras de surveillance installées dans les chambres des seniors qui résident dans des établissements d'hébergement des personnes âgées dépendantes. Les images reçues des caméras installées sont analysées localement en EHPAD 24H/24 et 7J/7 automatiquement. Lorsqu'une alerte est détectée, un message avec une image est envoyé vers un opérateur de la plate-forme d'assistance. La plate-forme d'assistance composée d'agents ou de ressources humaines disponibles 24H/24 et 7J/7, qui réceptionnent et traitent les alertes détectées et reçues.

Les Alertes sont à classer et à qualifier, par la suite par un agent, en 4 catégories:

- 1- Fausses Alertes: Pièces vides.
- 2- Fausses Alertes: Personnes actives.
- 3- Alertes niveau moyen: Personne assise.
- 4- Alertes niveau élevé: Personne couchée.

L'agent détermine si une action d'assistance est nécessaire ou non. En fait, les alertes avec un niveau moyen ou élevé nécessitent le contact par la ressource humaine de l'Établissement EHPAD par téléphone pour prévenir d'un incident survenu au numéro de la chambre concernée. Ce processus est décrit par la figure 1.

Le processus de gestion des incidents utilisé dans cette étude de cas est conforme au processus d'actions correctives / préventives ISO 9001. C'est un processus métier simple (comme montre la figure 1), mais il représente plusieurs contraintes fonctionnelles telles que : la scalabilité, une analyse de données en temps réel, l'obligation de maintenir des ressources (humaines et machines) limitées pour la viabilité de l'entreprise. Notre objectif est donc de proposer une approche qui permet de gérer les instances de ce processus selon un ordre de priorité, afin d'optimiser le temps de prise en charge de la personne en danger ainsi que les ressources humaines qui interviennent dans ce processus. Cette approche se base sur la détection des situations d'intérêts (simple ou complexe). Par exemple dans notre cas d'étude,



donner plus de priorité à un événement déclenché par un patient, si ce dernier a des besoins particuliers (fauteuil roulant, déambulateur, etc), ou par un patient récidiviste (chuteur répétitif).

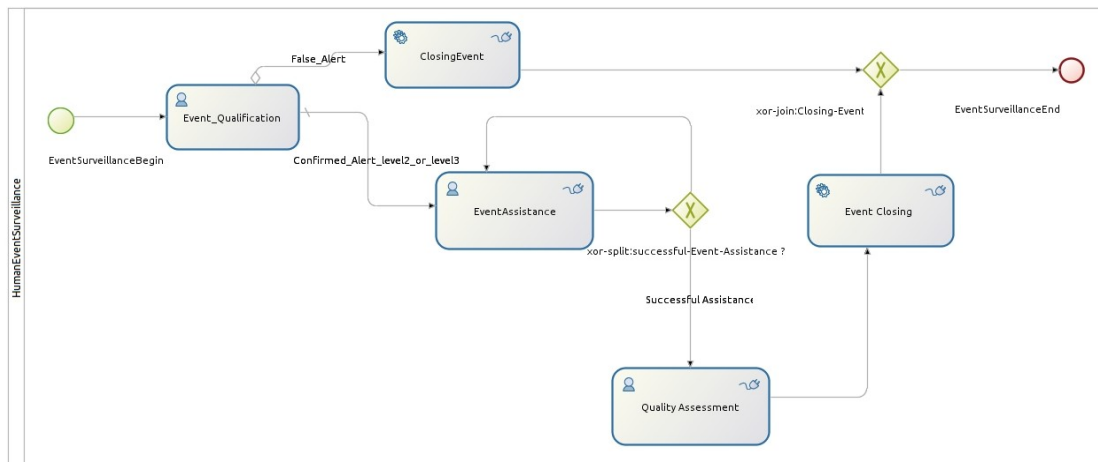


Figure 1. Processus de qualification et d'évaluation du niveau d'urgence des alertes

Dans la section suivante, nous allons présenter un état de l'art qui résume les différentes approches proposées dans la littérature, pour gérer l'ordonnancement des instances des processus métier.

### 3.État de l'art

Le « scheduling » traduit littéralement en français par l'ordonnancement, est défini comme un moyen de trouver une séquence d'exécution optimale et efficace des activités et des tâches d'un processus (métier), tout en respectant certaines contraintes telles que les contraintes de causalités, les contraintes temporelles ou les contraintes de ressources (humaines et/ou machines), et tout en évitant le sur-alimentation ou le sous-alimentation (Avanes et al., 2008). L'ordonnancement dans le BPM va de pair avec l'allocation des ressources, les deux ont un impact réciproque (Schulte et al., 2015). En fait, les approches d'allocation des ressources sont utilisées pour appuyer les décisions d'ordonnancement concernant l'affectation des tâches, des activités ou des instances aux ressources (humaines/machines) disponibles et convenables.

Les travaux de recherche portant sur ces deux notions, continuent à susciter un intérêt croissant de l'industrie et de la communauté scientifique. La nature des problèmes d'ordonnancement rencontrés par les organisations, oriente le choix des approches. En effet, ces problèmes diffèrent d'une organisation à l'autre, selon différents critères tels que : le domaine d'application, les objectifs visés (la performance, la qualité de services, etc), l'infrastructure dans laquelle s'exécutent

les processus métier en question (cloud computing, grid computing, etc), ou les métriques et les critères à optimiser (coût, temps de planification, temps d'exécution, etc). L'approche choisie dépend, en plus, du nombre de ces critères prédéfinis. Selon (Bessai et Charoy, 2016), les approches d'optimisation existantes peuvent être classées en deux grandes catégories : - Approche mono-critère, utilisé pour optimiser (minimiser ou maximiser) un seul critère. - Approche multicritère, où l'optimisation inclut plusieurs critères (qui peuvent être parfois conflictuels (Hofacker et al., 2001)), et le but est d'assurer une solution équilibrée et optimale sans compromettre un critère au détriment de l'autre.

Pour les contraintes temporelles, par exemple, différentes approches ont été proposées. Les auteurs dans (Eder et al., 2013) proposent une approche basée sur la notion d'échéance afin de respecter l'échéance globale du processus, tout en respectant toutes les autres contraintes temporelles externes. Ordonner l'exécution des activités dans les processus métier, peut améliorer considérablement les processus en question, comme montré par (Baggio et al., 2004), où les auteurs comparent plusieurs techniques connues d'ordonnement telles que : l'ordre aléatoire (Service In Random – SIRO), l'ordre selon la date d'échéance la plus proche (Earliest Due Date – EDD), ou encore l'ordre standard du premier entré premier sorti (First In First out – FIFO). En ce qui concerne les autres contraintes telles que la qualité de service (QoS) ou le coût, les approches d'ordonnement les plus utilisées sont les approches basées sur les méta-heuristiques, principalement les Algorithmes Génétiques qui sont fréquemment utilisés pour résoudre les problèmes d'ordonnement des tâches, des activités ou des instances dans les processus métier (Low et al., 2014), (Xu et al., 2016), (Ismaili-Alaoui et al., 2018a, 2018b).

Pour les contraintes de ressources, ce problème d'optimisation est plus compliqué, car il faut prendre en considération le type des ressources (humaines et/ou machines). Plusieurs approches ont été proposées dans la littérature pour résoudre ce problème. Les auteurs dans (van der Aalst et al., 2011) modélisent le problème d'optimisation de l'allocation des ressources dans le BPM comme un Processus de décision markovien, et ils utilisent un mécanisme d'allocation des ressources basé sur l'apprentissage par renforcement (reinforcement learning) pour prendre des décisions d'allocation en temps réel en minimisant les coûts à long terme et en améliorant ensuite la performance de l'exécution du processus métier. Alors que les auteurs dans (Huang et al., 2011) présentent une approche d'allocation des ressources basée sur l'exploration des règles (ou le rules mining) afin de découvrir des règles d'allocation intéressantes à partir du journal d'événements (event log).

La montée en puissance du cloud computing aide les organisations à gérer leurs activités de processus métier dans de meilleures conditions (haute performance et faibles coûts d'exploitation, etc.). Néanmoins, les approches proposées dans la littérature sont toujours confrontées à différents problèmes concernant la qualité de service, le coût et les contraintes de temps afin d'optimiser l'utilisation des ressources utilisées par le processus métier (Halima et al., 2017). Pour surmonter ce problème et ensuite profiter au maximum des différentes infrastructures qui facilitent l'exécution des processus métier et aussi pour optimiser l'utilisation de ces

ressources proposées ; de nouvelles approches s'orientent de plus en plus vers le BPM élastique en incluant l'élasticité dans les différents défis rencontrés dans le domaine du BPM tels que l'ordonnancement, l'allocation des ressources, etc, afin de surmonter les différents problèmes que les approches traditionnelles ne pouvaient pas résoudre (Schulte et al., 2015). Cependant, les organisations ne peuvent pas profiter pleinement des avantages du BPM élastique si la plupart des activités de leurs processus métier sont exécutées par des ressources humaines. En fait, les tâches et les activités dans les processus métier peuvent être exécutées par des ressources humaines ou des machines, car les processus métier sont différents des workflows scientifiques, car ils peuvent contenir des tâches automatiques et non automatiques. En général, les ressources humaines sont plus difficile à gérer car une ressource humaine peut exécuter d'autres tâches qui n'appartiennent pas au processus principal (Bessai, 2014), elles peuvent également n'être disponibles que pour des créneaux horaires spécifiques, et elles ont de nombreuses caractéristiques qui doivent être prises en considération telles que la disponibilité ou la fiabilité (Ismaili-Alaoui et al., 2018b).

L'élasticité est considérée comme un moyen réactif de faire face aux problèmes d'ordonnancement et d'allocation des ressources dans le BPM (surprovisionnement, sous-provisionnement). Ainsi, la proactivité représente une alternative à l'élasticité dans les problèmes d'ordonnancement des processus métier, en particulier lorsque ces processus intègrent des ressources humaines.

#### **4. L'approche proposée**

La gestion des processus métier basée sur les événements est principalement adaptée aux organisations qui ont des activités en temps réel impliquant des capteurs ou des dispositifs IoT qui collectent des données et génèrent de nouveaux événements, en surveillant leur environnement. Cependant, un système en temps réel doit avoir trois caractéristiques principales pour assurer un meilleur fonctionnement (Pielmeier et al., 2018) : 1) Haute disponibilité, 2) Faible latence et 3) Scalabilité horizontale. Et ces trois caractéristiques sont obligatoires pour réaliser un ordonnancement et une gestion des événements efficaces et en temps réel.

Afin d'assurer un ordonnancement basé sur la priorité en quasi-temps réel des instances d'un processus métier, nous avons recours au CEP, car il est considérée comme le standard pour l'analyse et la détection de situations en temps réel (Luckham, 2011).

Un événement, également appelé événement atomique, est un enregistrement instantané d'une activité dans un système (Luckham, 2011) à un moment donné, et il représente tout changement qui se produit ou se produira dans ce système. Alors qu'un événement complexe est un ensemble d'événements qui sont reliés entre eux par des opérateurs d'événements tels que l'agrégation, la causalité, la sémantique ou le temps (Robins, 2010). Le domaine de CEP vise à aborder plusieurs problèmes liés aux événements tels que le filtrage, le routage, la transformation et la détection d'événements complexes tout en traitant les événements atomiques et en utilisant des patterns d'événements prédéfinis (Darko, 2011). Ce genre de technologie facilite la

corrélation d'un très grand volume d'événements qui arrivent dans une période de temps limité, afin d'en extraire en temps réel des informations exploitables.

L'une des fonctionnalités offertes par les moteurs CEP est la détection de patterns d'événements, la Figure 2 (Boubeta-Puiget al., 2019) illustre cette fonctionnalité. Nous avons utilisé cette fonctionnalité pour détecter des incidents critiques (chutes de patients) dans notre processus.

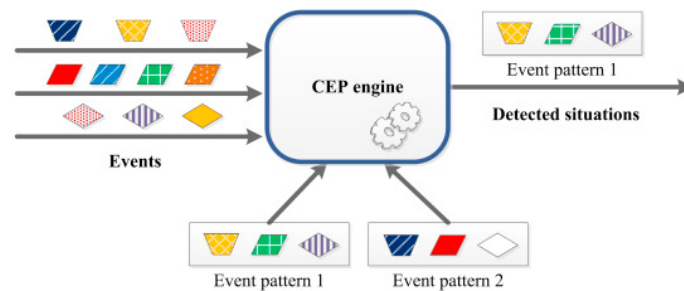


Figure 2. Détection des patterns d'événements

La détection des patterns se base principalement sur des règles. En nous basant sur le contexte de notre étude de cas, nous avons défini quelques règles qui nous aident à estimer le niveau de priorité de l'événement entrant. Pour définir ces règles, nous prenons en compte les informations disponibles sur le patient, ses incidents passés (chutes) et le cluster (Cluster de criticité) auquel ce patient (Source de l'événement) appartient. Nous avons essayé de définir manuellement certaines règles. Ainsi, nous sommes parvenus aux règles suivantes :

- **Si** la source de l'événement (le patient) appartient à un cluster de grande criticité **Alors** le nouvel événement généré par cette source pourrait être grave.
- **Si** le patient a des besoins particuliers (fauteuil roulant, déambulateur, etc.) **Alors** le nouvel événement généré par cette source pourrait être grave.
- **Si** le dernier événement généré par un patient dans un délai d'un mois était une alerte grave ou très grave, **Alors** le nouvel événement généré par ce patient (ou cette source) pourrait être grave.

La figure 3 illustre notre première tentative d'intégration du moteur CEP dans notre architecture IoT-BPM, pour la détection des patterns d'événements à base de priorité.

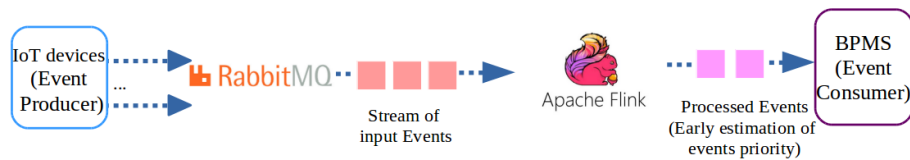


Figure 3. Détection des patterns d'événements

Les différents modules de cette approche fonctionnent comme suit :

- **Générateur d'événements** : dans cette approche, les événements sont généralement générés par des capteurs ou des dispositifs IoT qui surveillent leur environnement (caméras intelligentes).
- **Message broker** : Pour gérer les quantités d'événements reçus et devant être traités par le moteur CEP, nous utilisons un agent de messagerie qui assure la communication entre la source et la cible sur la base d'un mécanisme de publication/abonnement. Ce mécanisme asynchrone mis en œuvre par les messages brokers permet de découpler complètement les messages source et cible. En outre, les messages brokers peuvent également stocker les messages localement jusqu'à ce qu'ils puissent être traités par l'élément cible. Nous avons choisi RabbitMQ.
- **Moteur CEP** : Flink CEP est utilisé dans ce cas pour filtrer et traiter les événements entrants selon les règles prédéfinies, afin de détecter les événements ayant la plus haute priorité parmi le flux d'événements entrants.
- **Consommateur d'événements** : représente dans cette approche un système de gestion des processus métier (BPMS) où les processus sont gérés, exécutés et monitorisés.

Nous représentons dans la section suivante les résultats de l'implémentation de cette approche.

## 5. Résultats expérimentaux et discussions

Pour mettre en œuvre cette approche, nous avons utilisé FlinkCEP, qui est une bibliothèque de traitement d'événements complexes implémentée au-dessus de Flink<sup>1</sup>, utilisée pour détecter des patterns d'événements dans un flux d'événements, afin de ne retenir que ce qui est important dans le flux de données. Le choix de FlinkCEP a été basé sur plusieurs critères tels que la performance et le temps de réponse, le langage de programmation utilisé dans cette bibliothèque et enfin la communauté des programmeurs.

L'objectif de cette série d'expériences est de montrer l'intérêt d'intégrer le CEP dans la gestion des instances, déclenchées par des événements, selon un ordre de priorité. Nous avons implémenté deux solutions. La première solution (solution 1) sans le CEP et la deuxième (solution 2) avec le CEP. Afin de comparer les deux approches,

<sup>1</sup> <https://flink.apache.org/>

nous avons essayé de reproduire le même environnement expérimental. En effet, toutes nos expériences ont été réalisées sur un processeur Intel(R) Core (TM) i5-540 M 2.53 GHz. Et les deux approches ont été testées avec un ensemble de données provenant de notre étude de cas. Plus précisément, nous disposons d'un jeu de données de chutes de patients sur la période du 01-02-2016 au 12-06-2017, ce jeu de données est composé de 238228 observations générées par 81 patients : 89312 alertes sont de niveau 0 (faible), 148466 de niveau 1(moyen), 275 de niveau 2 (grave) et 175 de niveau 3. (grave) et 175 de niveau 3 (très grave).

Dans notre expérimentation pour les deux solutions, nous avons simulé plusieurs flux d'événements avec un nombre différent d'événements (générés à partir des événements historiques de notre ensemble de données). (100, 200, 300, 400, 500) pour l'accès non concurrent (voir Tableaux 1 et 2), et (200, 400, 600) pour l'accès concurrent (voir Tableaux 3 et 4). Notre objectif était de comparer l'évolution du temps de calcul des deux solutions en réponse à l'augmentation du nombre d'événements d'entrée (accès concurrent et accès non concurrent).

#### - Accès Non Concurrent

Tableau 1. Temps de calcul (sec) - Solution 1

Nombre total d'événements	Temps de calcul (sec)
100	19.0
200	19.45
300	25.48
400	31.3
500	37.4

Tableau 2. Temps de calcul (sec) - Solution 2

Nombre total d'événements	Temps de calcul (sec)
100	22.07
200	32.14
300	44.25
400	55.6
500	70.62

#### - Accès Concurrent

Tableau 3. Temps de calcul (sec) - Solution 1

Nombre total d'événements	Temps de calcul (sec)
200	48
400	72.1
600	108

Tableau 4. Temps de calcul (sec) - Solution 2

Nombre total d'événements	Temps de calcul (sec)
200	27
400	34
600	60.3

Comme nous pouvons le voir sur les figures 4 et 5, la solution 1 présente de meilleurs résultats que la solution 2 lorsque l'accès aux événements entrants n'est pas concurrent. Cependant, lorsque nous avons un accès concurrent, l'approche basée sur le CEP (solution 2) présente de meilleurs résultats, en particulier lorsque le nombre d'événements entrants augmente.

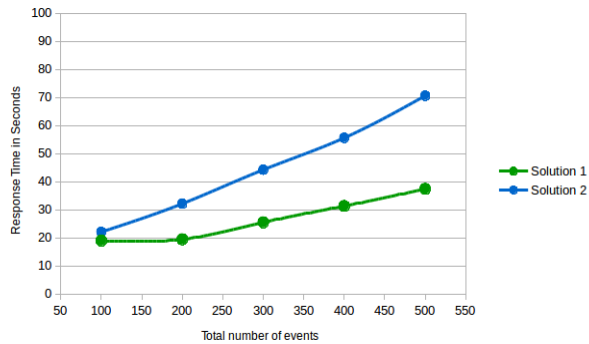


Figure 4: Événements entrants avec accès non concurrent

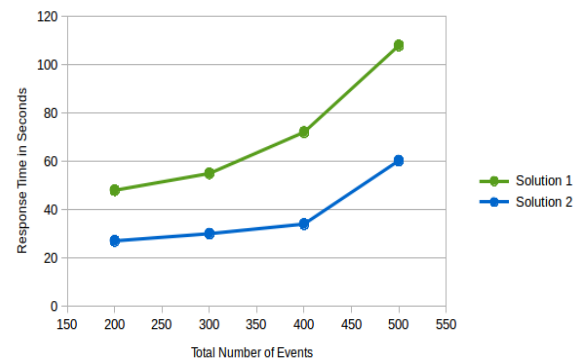


Figure 5: Événements entrants avec accès concurrent

Bien que la solution 1 semble être plus efficace pour de faibles volumes d'événements d'entrée, la solution CEP peut être plus performante, surtout si elle est mise en œuvre dans une architecture IoT et Big Data.

Pour les systèmes de gestion des incidents, il est très important de trouver un équilibre entre le traitement des événements en quasi-temps réel et la scalabilité afin d'obtenir un ordonnancement efficace et optimisé des instances de processus métier et une gestion efficace des événements. De plus, dans les cas réels, la plupart du temps, nous avons affaire à un accès concurrent des événements entrants.

Cela confirme donc l'efficacité de nos hypothèses selon lesquelles le CEP peut fournir de meilleurs résultats lorsqu'il est intégré dans une architecture IoT-BPM, et il peut également fournir de meilleurs résultats par rapport aux approches traditionnelles pour l'ordonnancement des instances des processus métier.

Les lecteurs intéressés peuvent consulter la solution complète que nous avons mise en œuvre à partir de GitHub<sup>2</sup>

## 6. Conclusion

Dans cet article, nous avons testé la possibilité d'utiliser le CEP afin de résoudre une partie des problèmes d'ordonnancement dans le BPM, en particulier l'exécution des instances de processus métier basées sur leur priorité. Notre approche proposée prouve que la gestion de la priorité des événements qui déclenchent les différentes instances d'un processus métier peut nous permettre de mieux planifier les instances, comparé aux approches traditionnelles. Les résultats obtenus prouvent que CEP doit être intégré, non seulement pour l'aspect contrôle et *monitoring*, mais également pour l'aspect planification et ordonnancement, en particulier pour les processus de

<sup>2</sup> [https://github.com/Abir-IA/CEPFlink\\_EventManagement](https://github.com/Abir-IA/CEPFlink_EventManagement)

gestion des incidents (incident management processes) ou les processus métier pilotés par les événements (event-driven business processes).

Notre approche présente certaines limites lorsqu'il s'agit de nouveaux événements qui peuvent ne correspondre à aucune règle de notre ensemble de règles prédéfinies. Dans ce cas, certains événements importants peuvent être considérés comme simples et moins prioritaires. C'est pourquoi, dans nos travaux futurs, nous prévoyons d'améliorer notre approche en utilisant des approches d'analyse prédictive combinées à un algorithme d'apprentissage automatique basé sur des règles (rule-based machine learning algorithm) afin de passer de la spécification manuelle des règles d'ordonnancement des instances de processus dans le moteur CEP à une spécification automatique, afin de réaliser une gestion proactive des incidents et une planification efficace des instances des processus métier.

### **Bibliographie**

- Anastassiou, M., Santoro, F. M., Recker, J., & Rosemann, M. (2016). The quest for organizational flexibility: driving changes in business processes through the identification of relevant context. *Business Process Management Journal*, 22(4), 763-790.
- Armbrust, Michael, et al. "A view of cloud computing." *Communications of the ACM* 53.4 (2010): 50-58.
- Avanes, A., & Freytag, J. C. (2008). Adaptive workflow scheduling under resource allocation constraints and network dynamics. *Proceedings of the VLDB Endowment*, 1(2), 1631-1637.
- Baggio, G., Wainer, J., & Ellis, C. (2004, March). Applying scheduling techniques to minimize the number of late jobs in workflow systems. In *Proceedings of the 2004 ACM symposium on Applied computing* (pp. 1396-1403). ACM.
- Baiyere, A., Salmela, H., & Tapanainen, T. (2020). Digital transformation and the new logics of business process management. *European Journal of Information Systems*, 29(3), 238-259.
- Bernard, C., Hallal, S., Nicolai, J. P., Montebourg, P. D. A., & Delaunay, M. (2013). *La Silver Économie, une opportunité de croissance pour la France*. Paris: CGSP.
- Bessai, K. (2014). *Gestion optimale de l'allocation des ressources pour l'exécution des processus dans le cadre du Cloud* (Doctoral dissertation, Université Paris1 Panthéon-Sorbonne).
- Bessai, K., & Charoy, F. (2016, November). Business process tasks-assignment and resource allocation in crowdsourcing context. In *2016 IEEE 2nd International Conference on Collaboration and Internet Computing (CIC)* (pp. 11-18). IEEE.
- Boubeta-Puig Juan, Gregorio Díaz, Hermenegilda Macià, Valentín Valero, and Guadalupe Ortiz. Medit4cep-cpn: An approach for complex event processing modeling by prioritized colored petri nets. *Information Systems*, 81:267-289, 2019.
- Cummins, F. A., (2009). *Building the Agile Enterprise*. Burlington: Elsevier.
- Darko Anicic, Paul Fodor, Sebastian Rudolph, Roland Stühmer, Nenad Stojanovic, and Rudi Studer. Etalis: Rule-based reasoning in event processing. In *Reasoning in event-based distributed systems*, pages 99-124. Springer, (2011).



- Eder, J., Panagos, E., & Rabinovich, M. (2013). Time constraints in workflow systems. In *Seminal Contributions to Information Systems Engineering* (pp. 191-205). Springer, Berlin, Heidelberg.
- Hofacker, I., & Vetschera, R. (2001). Algorithmical approaches to business process design. *Computers & Operations Research*, 28(13), 1253-1275.
- Huang, Z., van der Aalst, W. M., Lu, X., & Duan, H. (2011). Reinforcement learning based resource allocation in business process management. *Data & Knowledge Engineering*, 70(1), 127-145.
- Huang, Z., Lu, X., & Duan, H. (2011). Mining association rules to support resource allocation in business process management. *Expert Systems with Applications*, 38(8), 9483-9490.
- Halima, R. B., Kallel, S., Gaaloul, W., & Jmaiel, M. (2017, June). Optimal cost for time-aware cloud resource allocation in business process. In *2017 IEEE International Conference on Services Computing (SCC)* (pp. 314-321). IEEE.
- Ismaili-Alaoui, A., Benali, K., Baïna, K., & Baïna, J. (2018a, April). Business Process Instances Scheduling with Human Resources Based on Event Priority Determination. In *International Conference on Big Data, Cloud and Applications* (pp. 118-130). Springer, Cham.
- Ismaili-Alaoui, A., Baïna, K., Benali, K., & Baïna, J. (2018b, June). Towards Smart Incident Management Under Human Resource Constraints for an IoT-BPM Hybrid Architecture. In *International Conference on Web Services* (pp. 457-471). Springer, Cham.
- Koetter, F., & Kochanowski, M. (2015). A model-driven approach for event-based business process monitoring. *Information Systems and e-Business Management*, 13(1), 5-36.
- Lederer, M., Knapp, J., & Schott, P. (2017, March). The digital future has many names—How business process management drives the digital transformation. In *2017 6th International Conference on Industrial Technology and Management (ICITM)* (pp. 22-26). IEEE.
- Low, W. Z., De Weerd, J., Wynn, M. T., ter Hofstede, A. H., van der Aalst, W. M., & vanden Broucke, S. K. L. M. (2014, July). Perturbing event logs to identify cost reduction opportunities: A genetic algorithm-based approach. In *2014 IEEE Congress on Evolutionary Computation (CEC)* (pp. 2428-2435). IEEE.
- Luckham, D. C. (2011). *Event processing for business: organizing the real-time enterprise*. John Wiley & Sons.
- Mousheimish, R. (2017). *Combining the Internet of things, complex event processing, and time series classification for a proactive business process management* (Doctoral dissertation, Université Paris-Saclay).
- Poll, R., Polyvyanyy, A., Rosemann, M., Röglinger, M., & Rupperecht, L. (2018, September). Process forecasting: Towards proactive business process management. In *International Conference on Business Process Management* (pp. 496-512). Springer, Cham.
- Pielmeier, J., Braunreuther, S., & Reinhart, G. (2018). Approach for Defining Rules in the Context of Complex Event Processing. *Procedia CIRP*, 67, 8-12.
- Redlich, D., & Gilani, W. (2011, August). Event-driven process-centric performance prediction via simulation. In *International Conference on Business Process Management* (pp. 473-478). Springer, Berlin, Heidelberg.
- Reinartz, C., Metzger, A., & Pohl, K. (2015, June). Model-based verification of event-driven business processes. In *Proceedings of the 9th ACM International Conference on Distributed Event-Based Systems* (pp. 1-9). ACM.
- Robins, D. (2010, February). Complex event processing. In *Second International Workshop on Education Technology and Computer Science. Wuhan* (pp. 1-10).

- Schulte, S., Janiesch, C., Venugopal, S., Weber, I., & Hoenisch, P. (2015). Elastic Business Process Management: State of the art and open challenges for BPM in the cloud. *Future Generation Computer Systems*, 46, 36-50.
- Soffer, P., Hinze, A., Koschmider, A., Ziekow, H., Di Ciccio, C., Koldehofe, B., ... & Song, W. (2017). From event streams to process models and back: Challenges and opportunities. *Information Systems*.
- Weidlich, M., Ziekow, H., Mendling, J., Günther, O., Weske, M., & Desai, N. (2011, August). Event-based monitoring of process execution violations. In *International conference on business process management* (pp. 182-198). Springer, Berlin, Heidelberg.
- Van Der Aalst, W. M. (2013). Business process management: a comprehensive survey. *ISRN Software Engineering*, 2013.
- Xu, J., Liu, C., Zhao, X., Yongchareon, S., & Ding, Z. (2016). Resource management for business process scheduling in the presence of availability constraints. *ACM Transactions on Management Information Systems (TMIS)*, 7(3), 9.

---

# Analyse des Approches de Fouille d'Intentions : un Cadre de Comparaison

Rébecca Deneckère<sup>1</sup>, Elena Kornyshova<sup>2</sup>, Charlotte Hug<sup>1</sup>

<sup>1</sup>Centre de Recherche en Informatique (CRI),  
Université Paris 1 - Panthéon-Sorbonne, 12 Place de Panthéon, 75005, Paris, France  
[rebecca.deneckere@univ-paris1.fr](mailto:rebecca.deneckere@univ-paris1.fr)

<sup>2</sup> CEDRIC  
Conservatoire National des Arts et Métiers, Paris, France  
[elena.kornyshova@cnam.fr](mailto:elena.kornyshova@cnam.fr)

---

RÉFÉRENCE DE L'ARTICLE INTERNATIONAL. Cet article est un résumé de l'article :  
Rébecca Deneckère, Elena Kornyshova, Charlotte Hug: A Framework for Comparative  
Analysis of Intention Mining Approaches. RCIS 2021: 20-37

MOTS CLES : Intention ; Fouille d'intentions ; Cadre de Comparaison

KEYWORDS : Intention ; Intention Mining ; Comparative Framework

---

## 1. Introduction

(Merriam-Webster, 2020) définit une intention comme “une détermination à agir d’une certaine manière”. Il est dit en psychologie que notre schéma du sens commun considère les intentions comme un état d’esprit ; nos actions sont considérées comme intentionnelles (Bratman, 1987). Ces intentions peuvent être clairement explicites ou alors implicites, exprimées en langage naturel dans différentes sources de documents, des requêtes, des logs, etc. Beaucoup d’approches utilisent des techniques de fouille pour identifier les intentions des utilisateurs. Ce domaine de recherche est assez neuf et le terme de Fouille d’Intentions (IM) a différents sens selon la communauté. Les techniques proposées et leurs objectifs peuvent être très différents d’un domaine à l’autre. Nous conduisons une revue systématique de la littérature sur l’IM et notre première étape a été de construire un cadre de comparaison pour évaluer les différentes techniques les unes par rapport aux autres.

## 2. Cadre de Comparaison

Notre cadre de comparaison comporte quatre dimensions (figure 1). La dimension **Objet** s'intéresse à ce qu'est l'IM et donne la structure des intentions, leur taxonomie et leur formalisme. La dimension **Utilisation** s'interroge sur le pourquoi de l'utilisation de l'IM et caractérise l'approche sur l'objectif attendu, le domaine d'application, la source et la cible. La dimension **Méthode** indique les fondations théoriques de la méthode, si elle utilise de la classification ou des ontologies. La dimension **Outils** donne des détails sur la manière de mettre en œuvre la méthode. La distinction entre les dimensions Méthode et Outils n'est pas forcément la plus facile à comprendre mais nous pouvons la simplifier en utilisant une métaphore gastronomique : les méthodes correspondent aux recettes alors que les outils correspondent aux instruments de cuisine utilisés.

Cadre de comparaison des approches de fouille d'intentions			
Dimension Objet	Dimension Utilisation	Dimension Méthode	Dimension Outils
<ul style="list-style-type: none"> <li>• Structure</li> <li>• Taxonomies d'intentions</li> <li>• Formalisme de modélisation des intentions</li> </ul>	<ul style="list-style-type: none"> <li>• Objectif d'utilisation</li> <li>• Domaine d'application</li> <li>• Artefact Source</li> <li>• Artefact Cible</li> </ul>	<ul style="list-style-type: none"> <li>• Méthode d'apprentissage automatique</li> <li>• Méthode d'automatisation</li> <li>• Type d'enregistrement des observations</li> <li>• Méthode mathématique</li> <li>• Basée sur la classification</li> <li>• Basée sur les ontologies</li> <li>• Contexte</li> </ul>	<ul style="list-style-type: none"> <li>• Modèle mathématique</li> <li>• Type de classification</li> <li>• Nom de l'algorithme</li> <li>• Nom de l'ontologie</li> <li>• Support</li> <li>• Nom de l'outil</li> </ul>

Figure 1. Cadre de comparaison des approches de fouille d'intentions

## 3. Conclusion

Ce cadre est composé de quatre dimensions (objet, utilisation, méthode et outils). Il est utile (i) pour comparer la littérature existante sur l'IM et définir les différents challenges restants, (ii) pour un utilisateur d'IM, être capable de comparer rapidement les approches et de sélectionner la plus pertinente dans son cas et (iii) pour une nouvelle approche d'IM, se positionner par rapport à la littérature existante.

### Bibliographie

Merriam-Webster (accessed in november 2020) Definition of Intention, <http://www.merriam-webster.com/dictionary/intention>.

Bratman, M. (1987) Intention, plans, and practical reason. Harvard University Press.

---

# Anti-patrons d'alignement métier des SI

## Proposition de classification issue d'une expérience professionnelle

Jean-Philippe Gouigoux<sup>1</sup>, Dalila Tamzalit<sup>2</sup>

1. Directeur technique Groupe SALVIA Développement, Aubervilliers, France  
[jp.gouigoux@salviadeveloppement.com](mailto:jp.gouigoux@salviadeveloppement.com)

2. Nantes Université, CNRS, LS2N, F-44000 Nantes, France  
[dalila.tamzalit@univ-nantes.fr](mailto:dalila.tamzalit@univ-nantes.fr)

---

*RESUME.* Le présent document synthétise l'article présenté à ISD 2021 : Gouigoux, J. P., & Tamzalit, D. (2021, September). Business-IT alignment anti-patterns: a thought from an empirical point of view. In 29TH INTERNATIONAL CONFERENCE ON INFORMATION SYSTEMS DEVELOPMENT (ISD2021 VALENCIA, SPAIN).

*Mots-clés :* transformation digitale, alignement métier, gouvernance

---

### 1. Introduction et problématique

Des défauts récurrents d'alignement métier des SI ont été constatés sur une trentaine de SI audités par l'auteur industriel entre 2014 et 2019. Ce constat de récurrence, malgré la diversité des métiers, a amené à proposer une formalisation de ces défauts sous forme d'anti-patrons ainsi que des consignes pour les solutionner. L'objectif est de faciliter leur détection ainsi que leur correction par la communauté. Les défis sont nombreux : couplage fort, coûts des développements informatiques, taux élevé d'échecs des projets informatiques, lenteur d'adaptation aux changements réglementaires ou stratégiques des entreprises, etc. (Yeow et al., 2018), (Braun et al. 2007) (Luftman, 2004), (Luftman et al., 2007) (Hinkelmann et al., 2016). Notre principale question de recherche est : comment caractériser et modéliser ces erreurs d'alignement pour que leur détection et leur correction soient facilitées ?

### 2. Proposition

Les auteurs ont détaillé deux anti-patrons et présenté deux autres, selon un formalisme propre et en considérant quatre couches proches du modèle

d'urbanisation des SI du CIGREF<sup>1</sup>. Le premier anti-patron « *Pure technical integration* » correspond aux intégrations purement techniques, où des processus métier complexes, au lieu d'être implémentés par des enchaînements de fonctionnalités logicielles, sont implémentés à l'intérieur d'un seul logiciel. L'impact principal est alors un très haut degré de couplage et une forte difficulté à faire évoluer le processus. Un exemple de résolution de ce problème est ensuite décrit dans le cas des conseils régionaux par la mise en place d'un référentiel des tiers. Le deuxième anti-patron « *The functional silo dedicated IT subsystem* » concerne les sous-systèmes correspondant à des silos fonctionnels, dont les difficultés apparaissent quand le métier évolue et que des fonctionnalités se révèlent partagées. Le troisième, « *Monolith application* » est celui présenté par les applications monolithiques, dont la concentration en fonctionnalités rend périlleuse tout changement de version, voire menace la stabilité du SI. Le quatrième, « *Functional multiple implementations* » correspond à des implémentations logicielles multiples d'une même fonctionnalité, rendant la gestion automatisée des règles métier complexe car diffuse dans la technique.

### 3. Conclusion et perspectives

La plupart des travaux sur l'alignement métier/IT sont issus du management des SI. Seuls les travaux de (Brown et al., 1998) ont abordé les anti-patrons pour adopter une SOA. Conscients des limitations de ce premier article, les auteurs proposent des pistes futures comme la généralisation des propositions et l'automatisation de la détection de 10 autres anti-patrons.

#### Principales références

- (Yeow et al., 2018) Yeow, A., Soh, C., & Hansen, R. Aligning with new digital strategy: A dynamic capabilities approach. *The Journal of Strategic Information Systems*, 27(1), 43-58.
- (Braun et al. 2007) Braun, C., & Winter, R. Integration of IT service management into enterprise architecture. In *Proceedings of the 2007 ACM symposium on Applied computing* (pp. 1215-1219).
- (Luftman, 2004) Luftman, J. (2004). Assessing business-IT alignment maturity. *Strategies for information technology governance*, 4, 99.
- (Luftman et al., 2007) Luftman, J., & Kempaiah, R. (2007). An Update on Business-IT Alignment: "A Line" Has Been Drawn. *MIS Quarterly Executive*, 6(3).
- (Hinkelmann et al., 2016), Hinkelmann, K., Gerber, A., Karagiannis, D., Thoenssen, B., Van der Merwe, A., & Woitsch, R. (2016). A new paradigm for the continuous alignment of business and IT: Combining EA modelling and enterprise ontology. *Comp. in Industry*, 79, 77-86.
- (Brown et al., 1998) Brown, W. H., Malveau, R. C., McCormick, H. W., & Mowbray, T. J. (1998). *AntiPatterns: refactoring software, architectures, and projects in crisis*. JW & Sons, Inc.

---

<sup>1</sup> <https://www.cigref.fr/>

---

# Vers un modèle d'équilibre de maturité numérique pour les organisations publiques

**Mateja Nerima, Jolita Ralyté**

*ISS, CUI, Université de Genève  
Battelle bâtiment A, 7 Route de Drize, 1227 Carouge, Suisse  
jolita.ralate@unige.ch*

---

*REFERENCE DE L'ARTICLE INTERNATIONAL. Cet article est un résumé de l'article :  
Mateja Nerima, Jolita Ralyté: Towards a Digital Maturity Balance Model for Public  
Organizations. RCIS 2021: 295-310, LNBIP, vol 415, Springer 2021.*

*MOTS-CLÉS : Transformation digitale, modèle de maturité numérique, organisation publique,  
équilibre de maturité numérique.*

*KEYWORDS: Digital Transformation, Digital Maturity Model, Public Organization, Digital  
Maturity Balance*

---

## 1. Contexte et objectifs

Le phénomène de la transformation digitale touche actuellement presque tous les secteurs d'activité. Selon Westerman et al. (2014), c'est un processus nécessaire pour atteindre la maîtrise numérique, qui conduit à de meilleurs profits, production et performance. Les organisations privées et publiques sont confrontées au défi de la croissance rapide de la digitalisation, mais leurs capacités de conduire la transformation ne sont pas toujours au point. Mesurer la maturité numérique d'une organisation est une étape cruciale dans le processus de sa digitalisation. Ceci permet de déterminer et hiérarchiser les objectifs de transformation, et estimer les moyens et ressources nécessaires pour les atteindre (Pöoppelbuß and Röglinger, 2011). Les caractéristiques et les enjeux de la transformation numérique sont propres à chaque secteur d'activité voire à chaque type d'organisation (Kane *et al.*, 2017). Par conséquent, chacun d'eux peut nécessiter un modèle spécifique de maturité numérique.

Dans ce travail, nous avons accordé une attention particulière au secteur public qui, selon notre analyse de l'art dans le domaine, n'a pas de modèle de maturité numérique qui lui serait dédié spécifiquement et les modèles existants ne sont que peu adaptés. Par conséquent, l'objectif de ce travail a été de développer un modèle de maturité numérique pour les organisations publiques. Pour tenir compte de la diversité des organisations publiques en termes d'activité et de taille, nous avons

développé un modèle *d'équilibre de maturité numérique* dans lequel chaque dimension de maturité est évaluée en tenant compte du ratio d'importance de cette dimension dans l'organisation.

## 2. Développement du modèle

Pour développer notre modèle d'équilibre de maturité numérique pour les organisations publiques, nous avons suivi une approche exploratoire en quatre étapes :

*Étape 1 – étude de l'état de l'art* : Nous avons sélectionné et étudié 20 modèles de maturité numérique existants afin d'extraire les critères les plus répandus et les plus pertinents pour les organisations publiques.

*Étape 2 – présélection et validation* : La première mouture de notre modèle de maturité numérique comportait 46 critères classés en 14 catégories. Afin d'inclure les utilisateurs potentiels du modèle dans le processus de son développement, nous avons mené une enquête en ligne auprès de 50 organisations publiques suisses dont 15 ont participé et répondu au questionnaire. L'enquête a permis de valider certains critères et d'en identifier d'autres. Par ailleurs, elle a confirmé notre perception que la maturité numérique n'a pas forcément le même sens d'une organisation à l'autre et nous a donné l'idée de travailler la notion d'équilibre de la maturité numérique plutôt que l'évaluation directe.

*Étape 3 – développement du modèle et de l'outil d'évaluation* : Plusieurs itérations ont été nécessaires pour construire le modèle d'équilibre de maturité numérique en deux axes : d'un côté les critères de maturité organisés en 5 dimensions (données, gouvernance TI, stratégie, organisation et processus) et de l'autre côté les métriques pour mesurer le ratio d'importance de chaque dimension de maturité dans l'organisation. Un outil d'auto-évaluation combinant les deux axes a été également développé.

*Étape 4 – évaluation* : Les 15 organisations participantes de la première enquête ont été invitées à tester l'outil d'auto-évaluation et à remplir un questionnaire. 7 d'entre elles ont accepté de participer. Dans l'ensemble, le modèle et l'outil d'auto-évaluation ont été jugés utiles et pertinents, mais le développement des deux nécessite des itérations supplémentaires.

## Bibliographie

- Kane G.C., Palmer D., Phillips A.N., Kiron D., Buckley N. (2017). *Achieving Digital Maturity: Adapting Your Company to a Changing World*. Deloitte University Press.
- Pöoppelbuß J., Röglinger, M. (2011). What makes a useful maturity model? A framework of general design principles for maturity models and its demonstration in business process management. *Proceedings of ECIS 2011*, 28, AIS eLibrary.
- Westerman G., Bonnet D., McAfee A. (2014). *Leading Digital – Turning Technology into Business Transformation*. Harvard Business Review Press. Boston, Massachusetts.



# Méta-modèle des concepts et processus d'analyse des risques selon les normes de cybersécurité

Christophe Ponsard<sup>1</sup>, Valery Ramon<sup>1</sup>, Mounir Touzani<sup>2</sup>

1. CETIC - Centre de recherche, Gosselies, Belgique  
*{christophe.ponsard, valery.ramon}@cetic.be*

2. Chercheur indépendant, Toulouse, France  
*mounir.touzani@inrae.fr*

---

*RÉSUMÉ. De nos jours, les systèmes d'information d'entreprise et de contrôle industriel sont fortement exposés aux menaces de cybersécurité. Dans de nombreux domaines, des mesures de protection sont activement déployées et encadrées par des normes ou standards. Il est fondamental de s'assurer que les risques de cybersécurité sont bien identifiés et contrôlés sur la base d'une analyse des menaces et d'une évaluation des risques. Cet article passe en revue les principales normes et permet de comparer des variantes dans différents domaines (systèmes d'information, systèmes industriels, automobile et aéronautique) afin de dégager un méta-modèle commun capturant tous les concepts manipulés durant une analyse des risques et le processus suivi pour la réalisation, éventuellement par itération et affinement. Enfin, nous illustrons et discutons son application sur un cas d'utilisation pour réaliser une analyse des risques dirigée par les modèles.*

*ABSTRACT. Today, corporate information systems and industrial control systems are highly exposed to cybersecurity threats. In many areas protective measures are actively deployed and supported by standards. A common foundation is to ensure that cybersecurity risks are properly identified and controlled based on a threat analysis and a risk assessment. This paper reviews and compares the different variants of standards in key domains (information systems, industrial systems, automotive and aeronautics) in order to derive a common meta-model capturing all the concepts required for a precise risk analysis as well as the process followed to perform it, possibly through iteration and refinement. We then illustrate and discuss how to deploy it for a model-driven risk analysis process.*

*MOTS-CLÉS : Cybersécurité, analyse de risques, modélisation, ISO 27000, EBIOS, IEC 62443, SAE 21434, DO-326, outillage*

*KEYWORDS: Cyber security, risk analysis, modelling, ISO 27000, EBIOS, IEC 62443, SAE 21434, DO-326, tool support*

---

## 1. Introduction

De nombreux secteurs d'activité sont devenus tributaires de systèmes d'information ou de contrôles industriels pour leur fonctionnement quotidien. Si cela contribue à les rendre plus réactifs, automatisés et compétitifs, la quantité de plus en plus importante de logiciels, combinée à un haut degré de complexité des systèmes et de leurs interconnexions accroît leur exposition aux menaces de cybersécurité. De récents rapports sur les menaces confirment les principaux fils conducteurs (logiciels malveillants, attaques en ligne, phishing) et leur évolution vers des attaques plus fréquentes et plus ciblées (ENISA, 2020). Les technologies numériques s'étant installées au cœur de toutes nos infrastructures critiques, en assurer la cybersécurité devient une préoccupation essentielle dans de nombreux secteurs industriels afin de les garder opérationnelles, voire d'en préserver la sûreté de fonctionnement. De plus, l'Europe a pris des actions en la matière pour imposer la directive NIS (Network Information System) dans une série de domaines de services essentiels (transport, énergie, traitement de l'eau, etc.) (EU, 2016), d'ailleurs en voie d'élargissement (services postaux, secteur alimentaire, fabrication des dispositifs médicaux...).

Les exigences principales de chaque secteur peuvent être remplies en mettant en place des mesures de protection capables d'assurer des propriétés clés de ses services généralement en termes de disponibilité, intégrité et confidentialité. Pour ce dernier point, lorsque des données personnelles sont concernées, les exigences sont renforcées par le RGPD (Règlement Général de Protection des Données) (European Commission, 2016). De telles mesures se déclinent sur plusieurs lignes de défense (protection, réaction, récupération) et s'appuient sur une démarche de gestion des risques qui doit au préalable identifier les risques, les évaluer avant de décider des mesures qui permettent de les traiter, par exemple, en les réduisant à un niveau acceptable. Pour atteindre ces objectifs, cette démarche doit englober toutes les activités coordonnées et nécessaires afin de diriger et contrôler une organisation en matière de risque (ISO, 2009).

La mise en place d'une telle démarche requiert des normes dont le degré d'aboutissement et d'adoption varie selon les domaines. Ainsi, si le secteur des technologies de l'information (TI) dispose de la série intégrée des normes ISO27K depuis 2005 (ISO, 2013), d'autres secteurs ont été plus tardifs à se doter des normes : les systèmes industriels et l'aéronautique vers 2010 et l'automobile en 2021. Cette évolution est positive et cette adoption par domaine est adéquate car chaque domaine a ses spécificités en termes de biens à protéger et technologies mises en oeuvre, de type IT (Information Technology) et/ou OT (Operation Technology) (BSI, 2020). Cependant, la diversité des normes peut aussi constituer un obstacle à leur mise en oeuvre notamment quand il s'agit d'en considérer plusieurs ou encore par rapport à la disponibilité de plateformes outillées soutenant efficacement le travail d'analyse des risques.

Notre objectif consiste à dégager un socle commun suffisamment riche pour permettre la mise en place d'approches efficaces basées sur des modèles et assurer un bon niveau de précision de l'analyse, plutôt que des approches documentaires qualitatives encore fréquemment de mise. Cela est envisageable car toutes les méthodes partagent

de nombreux concepts et processus issus d'une norme générale de gestion des risques (ISO, 2009), s'appuyant sur des notations et outils similaires comme des diagrammes de contexte, (mis)use cases, des descriptions d'infrastructures et parfois des arbres de menaces. Ceux-ci étant mis en oeuvre en fonction des spécificités d'un secteur et du degré de maturité de l'organisation. La contribution centrale est de réaliser un méta-modèle riche en concepts et processus qui permettent de capturer, comprendre et manipuler les analyses des risques prescrites par les différentes normes. Notre approche procède par enrichissement plutôt que par abstraction : nous essayons, d'une part, de mettre en évidence le socle commun, d'autre part, de capturer des variantes plus riches qui ajoutent de la précision, l'expressivité et la modularité dans la démarche. Ceci contraste avec d'autres approches déjà réalisées, ce qui nous a inspiré en partie.

La structure de cet article reflète la méthodologie suivie. La section 2 passe en revue les diverses normes concernées et en fait une synthèse dans quatre secteurs clefs. La section 3 dégage un méta-modèle qui donne une vision globale en termes de concepts manipulés : notions de biens, risques, menaces, mesures... et d'un processus de conduite d'une analyse des risques qui peut être plus ou moins élaborée selon les cas. Ensuite, la section 4 illustre et discute la mise en oeuvre de ces abstractions sur plusieurs scénarios. Enfin, nous concluons et présentons nos travaux futurs.

## 2. État de l'art

Cette section passe en revue des processus d'analyse des risques prescrits par plusieurs normes. Après la présentation du cadre générique de l'ISO 31000, nous décrivons des normes spécifiques à la (cyber) sécurité dans quatre secteurs clefs : les systèmes d'information purs (ISO27K), les systèmes industriels mixant de l'IT et de l'OT (IEC 63443), l'automobile (SAE 21434) et l'aéronautique (DO-236). Nous nous concentrons ici uniquement sur des normes protégeant des organisations et non pas des produits spécifiques qui sont pris en charge par d'autres types de normes, notamment les critères communs.

### 2.1. Cadre de gestion des risques : ISO 31000

Un processus générique d'analyse des risques est spécifié par la norme ISO 31000 (ISO, 2018) qui structure les autres normes présentées. Voici les principales étapes identifiables dans la Figure 1:

- **identification des risques**, identifie, permet de comprendre et décrire les risques. Elle prend en compte à la fois ce qui peut entraver mais aussi aider à l'atteinte des objectifs.
- **analyse des risques**, vise à comprendre la nature des risques et leur caractéristiques : type d'incertitude, sources, conséquences, probabilité, scénarios, contrôles existants et leur efficacité. Les risques sont souvent difficiles à quantifier avec précision. Une approche qualitative est souvent adoptée.

– **évaluation des risques**, soutient le processus de décision selon les critères de l'entreprise pour décider de mesures complémentaires. Elle s'appuie sur l'analyse des risques qui est souvent résumée à l'aide d'une matrice des risques telle que celle de la figure 3. Les risques nécessitant des mesures sont identifiés et priorisés.

– **traitement des risques**, sélectionne et met en œuvre les actions résultant du processus de décision. Il doit également prendre en compte les aspects de coût, d'efforts et de délais. Pour chaque risque, les options possibles sont : ne rien faire (accepter), envisager des actions supplémentaires (atténuation), partager les conséquences (transfert), supprimer la source (évitement). Comme il n'est pas réaliste d'éliminer totalement les risques et que les actions peuvent en introduire de nouveaux, le degré de risque résiduel doit être évalué. Un plan d'actions est alors préparé et exécuté.

– **surveillance des risques**, le processus suit une boucle globale avec des réévaluations régulières (complètes ou incrémentales), suite à l'évolution de l'organisation et de son exposition aux risques.

## 2.2. Systèmes d'information : cadre ISO 27005

La norme ISO 27000 (ou « ISO27K » en abrégé) est une famille des normes relatives au déploiement d'un système de gestion des risques de sécurité de l'information. Elle définit le vocabulaire (27000), les exigences du système de gestion (27001), les contrôles (27002) et une approche de gestion orientée sur le risque (27005).

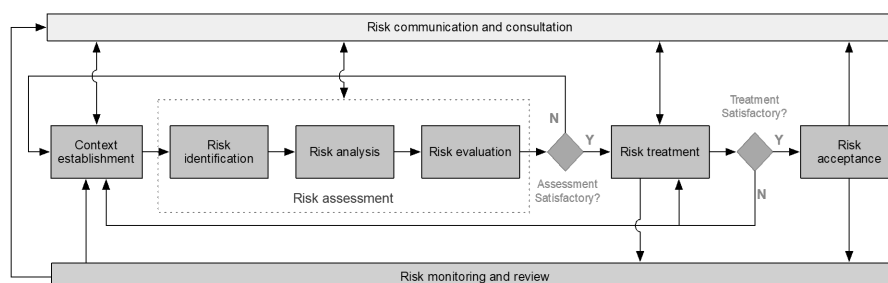


FIGURE 1. ISO 27005 Analyse des menaces et évaluation des risques

Le processus d'analyse des risques de sécurité de l'information est une spécialisation de la norme ISO 31000, comme le montre la figure 1. La norme ISO 27005 n'impose pas de méthodologie, mais un ensemble d'exigences à atteindre :

- **Établissement du contexte**, définit la portée, les limites, rôles, responsabilités.
- **Identification des risques**, explicite les principaux actifs à protéger et les actifs de soutien (logiciels, matériels, infrastructure physique, personnel...), de même que les sources de menaces (internes, externes) et les contrôles existants.
- **Estimation des risques**, fournit une estimation, généralement qualitative, de chaque risque en termes de probabilité et de conséquence en fonction de l'impact sur les propriétés de confidentialité, d'intégrité ou de disponibilité.

- **Évaluation des risques**, produit une liste des risques classés par ordre de priorité selon les critères d'évaluation des risques de sécurité.

Le processus comprend deux portes de décision explicites contrôlant respectivement le niveau de qualité de l'évaluation des risques et du plan de traitement des risques. De nombreuses implémentations de la norme ISO27005 sont disponibles, par exemple, EBIOS , MEHARI ou OCTAVE. Dans le cadre de cet article, nous nous limitons à EBIOS (ANSSI, 2010).

### 2.3. Systèmes d'information : EBIOS

EBIOS (Expression des Besoins et Identification des Objectifs de Sécurité) est une méthode française développée par la communauté EBIOS et soutenue par l'ANSSI, l'autorité française de défense des systèmes d'information. Elle est conforme à la norme ISO27005. Malgré l'évolution vers un schéma plus agile, nous prenons en compte la version 2010 dont la structure est représentée dans la Figure 2.

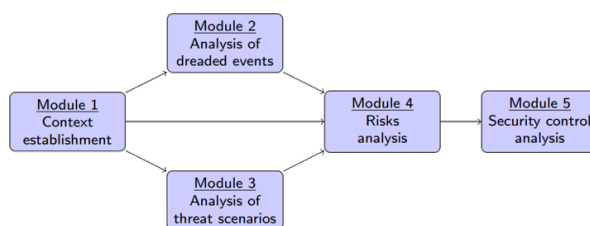


FIGURE 2. Activités EBIOS

Globalement, voici la correspondance avec la norme ISO 27005 et les spécificités de mise en œuvre :

- **L'établissement du contexte** nécessite d'énoncer les objectifs de l'organisation, le périmètre considéré, les échelles qualitatives de mesure de la confidentialité, la disponibilité et l'intégrité. Cela couvre aussi l'identification des actifs « primaires » (données et processus métier) et « secondaires » (infrastructure informatique et personnes). Le niveau de complexité de l'attaque est également identifié.

- **L'analyse des événements redoutés** est la première partie de l'estimation des risques. Il s'agit d'une approche descendante axée sur l'impact au niveau du métier. Elle estime les conséquences de la perte de confidentialité, d'intégrité et de disponibilité sur les différents actifs primaires. Les sources de menaces sont aussi identifiées.

- **L'analyse des scénarios de menace** est la deuxième partie de l'estimation des risques, réalisée en parallèle avec l'analyse des événements redoutés. Elle fonctionne de manière ascendante en considérant les scénarios de menace affectant les ressources de support. La probabilité est estimée sur une échelle qualitative contextuelle, avant et après l'application des mesures existantes.

- **L'analyse des risques** combine les résultats des deux étapes précédentes pour estimer chaque risque et produire une matrice des risques telle que décrite dans la

figure 3. Le processus peut combiner plusieurs scénarios (cas le plus défavorable). Une hiérarchisation est effectuée et le type d'action est décidé parmi les options proposées dans la norme ISO 31000 (éviter, accepter, atténuer, transférer).

– **L'analyse des contrôles de sécurité** détermine les mesures nécessaires pour traiter les risques. Ces contrôles peuvent être organisés sur plusieurs lignes de défense de prévention, de protection et de récupération. Les conseils sont fournis à l'aide d'une base de connaissances en lien avec la liste des contrôles ISO 27002.

<b>Gravité</b>	4. Critique		Indisponibilité des alarmes	Perte d'intégrité des bilans	
	3. Importante		Perte d'intégrité des données Indisponibilité des alarmes Indisponibilité des bilans		
	2. Limitée		Indisponibilité des données		
	1. Négligeable				
		1. Minimale	2. Significative	3. Forte	4. Maximale
<b>Vraisemblance</b>					

FIGURE 3. *Matrice des risques*

#### 2.4. *Système industriels : IEC/ISA 62443*

La norme CEI/ISA 62443 a été développée par les comités ISA99 et CEI pour améliorer la sécurité des composants ou des systèmes utilisés dans l'automatisation et le contrôle industriel (IACS) (IEC, 2010a). Elle est le double de la norme CEI 61508 (IEC, 2010b) qui vise la sûreté de fonctionnement. Son périmètre est plus large que la norme ISO27K car il couvre non seulement l'IT mais aussi l'OT pour assurer le contrôle et la supervision de systèmes industriels ou de transport (par exemple ferroviaire). Elle est divisée en plusieurs parties couvrant les généralités, les politiques, les systèmes et les composants. L'évaluation des risques de sécurité (SRA) est couverte par la partie 62443-3-2. Sa classification des exigences de sécurité est également plus riche et avec 7 catégories qui couvrent les classiques : intégrité (FR3), confidentialité (FR4) et disponibilité (FR7) mais aussi 4 autres caractéristiques : le contrôle d'identification et d'authentification (FR1) de tous les utilisateurs. Puis, le contrôle de l'utilisation pour s'assurer du respect des privilèges requis (FR2), la restriction des flux sur les zones et conduits (FR5) ainsi que la réaction aux violations de sécurité (FR6). Elle introduit aussi quatre niveaux de sécurité (SL) afin de classifier et raisonner sur les risques en fonction de la criticité avec une progression en termes d'intentionnalité, sophistication et ressources mises en oeuvre.

Voici les étapes de l'évaluation des risques de sécurité (SRA) :

- ZCR1 : identification du système (cf. établissement du contexte dans l'ISO27K).
- ZCR2 : analyse des risques de haut niveau pour identifier les risques de cybersécurité liés aux opérations critiques. Elle peut se baser sur une analyse HAZOP.
- ZCR3 : partition en zones et conduits. Il s'agit d'une forme plus spécialisée de modélisation du système afin de fournir une isolation et organiser les lignes de défense.

- ZCR4 : évaluation de l’acceptabilité du risque au vu des mesures existantes.
- ZCR5 : en cas de risque non acceptable, une évaluation détaillée est effectuée au niveau des zones et des conduits. Des mesures supplémentaires sont identifiées et ZCR4 est réévaluée jusqu’à atteindre un niveau acceptable.
- ZCR6 : documentation des exigences/hypothèses/contraintes de cybersécurité.
- ZCR7 : approbation du rapport d’analyse par le propriétaire du bien.

### 2.5. Automobile : ISO 21434

Cette norme récente (2021), vise à établir un consensus sur les principaux problèmes de cybersécurité dans le domaine automobile (ISO, 2020). Elle remplace les bonnes pratiques J3061 (SAE, 2016) par des recommandations plus structurées. Son champ d’application concerne les véhicules routiers (voitures, camions, bus) et couvre leurs sous-systèmes, composants, connexions et logiciels. Son objectif est de garantir que les constructeurs et tous les participants de la chaîne d’approvisionnement disposent de processus structurés dès la phase de conception.

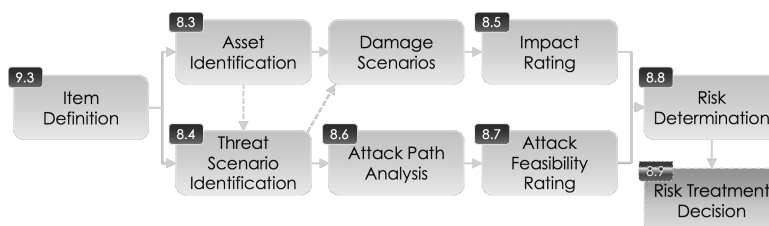


FIGURE 4. ISO 21434 Analyse des menaces et évaluation des risques

La norme est structurée en 10 sections et 15 clauses. Elle commence par définir (1) le champ d’application, (2) la référence normative, (3) le glossaire, (4) les considérations générales et (5/cloauses 5-6-7) l’approche de gestion. Ensuite, (6/cloause 8) se concentre sur l’évaluation des risques. Elle est suivie de trois sections couvrant respectivement (7/cloause 9) la phase de conception, (8/cloauses 10-11) le développement du produit et (9/cloauses 12-13-14) le produit, l’exploitation et la maintenance. La dernière section (10/cloause 15) traite des processus de soutien.

La norme n’impose aucune méthode d’évaluation des risques mais sa clause 8 rappelle les étapes requises, dans le même esprit que l’ISO 27005. La figure 4 détaille ses activités qui suivent un processus composé d’une branche métier identifiant les actifs (8.3) et évaluant les impacts (8.5), ainsi qu’une branche technique se concentrant sur l’analyse des menaces (8.4), l’identification du chemin d’attaque (8.6) et l’évaluation de la faisabilité (8.7). Les deux branches sont utilisées pour déterminer (8.8) et traiter les risques (8.9). Elle introduit aussi des niveaux d’assurance sur 4 niveaux CAL (Cybersecurity Assurance Level) totalisant 27 activités classées en 7 catégories. La norme ne décrit pas de technologies ou de solutions spécifiques et ne donne pas de recommandations sur les contre-mesures.

## 2.6. Aéronautique : DO-326/ED202

Si l'aéronautique disposait depuis longtemps de normes strictes en matière de sûreté de fonctionnement (DO-178), elle ne s'est dotée de normes de cybersécurité que vers 2010 avec la DO-326/ED-202, largement inspirée de l'ISO 27K et du standard de facto plus général SAE ARP 4754 (David, 2019). Elle aborde les considérations essentielles/de guidance et est complétée par d'autres documents pour les aspects services, systèmes sols et de plus haut niveau qui relèvent aussi des responsabilités externes (équipementiers). Elle ne spécifie aucune mesure de sécurité ou technique ni méthode mais donne plutôt des stratégies/tactiques applicables au domaine et recommande l'utilisation d'autres normes telle que l'ISO27K pour la mise en oeuvre.

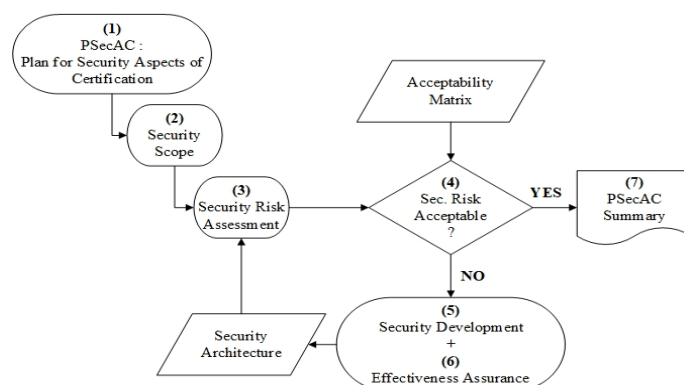


FIGURE 5. Analyse des risques selon la norme aéronautique DO-326

Cette norme détaille un processus de sécurité de la navigation aérienne (en anglais Airworthiness Security Process - AWSP) qui comporte 7 activités dont une analyse des risques qui comprend des étapes de définition de périmètre, d'identification des menaces, de leur caractérisation et évaluation de leur niveau. Les activités sont plus précisément représentées dans la figure 5, en parallèle avec les activités relatives au développement et à la sûreté de fonctionnement. On note la décomposition en trois étapes : préliminaire, système, spécifique à l'appareil lui-même. Les activités d'assurance sont réparties sur 4 niveaux SAL (Security Assurance Level) et sont au nombre de 118 classées en 13 sections.

## 2.7. Synthèse

La table 1 donne une synthèse comparative des différentes normes présentées en résumant leur domaine, le périmètre, le type d'exigence de sécurité, éventuellement le mécanisme de décomposition, le processus associé, la présence de listes d'exigences ou de contrôles ainsi que la présence de niveaux de sécurité/maturité (notamment en référence au modèle CMMI - Capability Maturity Model Integration).

Globalement, on constate plusieurs convergences qui permettent d'envisager la suite du travail de rapprochement des concepts et processus. EBIOS (ISO 27005) ap-



TABLEAU 1. *Synthèse comparative des différentes normes*

Sujet	EBIOS (ISO27K)	IEC 62443	ISO 21434	DO-326
Domaine	Systèmes d'information	Systèmes industriels	Automobile	Aéronautique
Périmètre	IT	IT/OT	IT/OT	IT/OT
Exigences clefs	3 (CID)	7 (FR1-FR7)	3 (CID)	3 (CID)
Décomposition	non	oui préliminaire puis détaillée (zones/conduits)	notion de sous-système et de chemin d'attaque	oui (préliminaire, système, bord)
Processus	ISO 31000 avec branche métier et infra	ISO 31000 avec niveau global et niveau détaillé	ISO 31000 avec branche métier et infra	ISO31000 (via inspiration ISO27005)
Liste d'exigences ou contrôles	basés sur l'ISO 27002 2013: 14 chap., 114 mesures 2022: 4 chap., 93 mesures	8 catégories (7 FR+1) 72 exigences	pas de liste	via norme externe
Niveaux de sécurité	Prioritisation libre	Niveaux SL0 à SL4 (cible/réalisé)	–	–
Niveau d'assurance	hors scope	4 niveaux de maturité basé sur CMMI 8 pratiques, 45 exigences	CAL 1 à 4 (informative) 7 catégories, 27 activités	SAL 0 à 3 13 sections, 29 exigences, 118 activités

paraît plus léger et moins capable de traiter des systèmes par approche de décomposition. IEC62443 est très complète. L'ISO 21434 en automobile reste assez monolithique mais plus détaillée qu'EBIOS, tandis que la DO-326 est structurée mais est essentiellement une norme chapeau qui défère sa mise en oeuvre à d'autres.

### 3. Méta-modèle des concepts et processus de gestion des risques

Cette section décrit un référentiel de gestion des risques indépendamment de normes spécifiques sous la forme de méta-modèles couvrant à la fois les concepts manipulés par ces normes mais aussi le processus de conduite de l'analyse des risques. Avant de décrire ces méta-modèles, le paragraphe suivant présente la finalité et la démarche.

#### 3.1. Finalité et démarche de construction du méta-modèle

La synthèse des normes a montré de fortes similitudes notamment dans la démarche sous-jacente de gestion des risques. On constate bien sûr des différences mais qui relèvent plus de la variante ou du raffinement. Ceci permet d'envisager des bases communes de conceptualisation. Le point fondamental est de permettre une démarche explicitement orientée modèle en évolution de démarche basée sur des processus documentaires. Plus précisément, la modélisation explicite des concepts et des processus permettant de :

- disposer d'un référentiel unique pour ensuite appréhender un domaine plus spécifique, par exemple, l'automobile ou un système industriel de contrôle.
- faciliter le processus de passage à une autre norme, à partir d'un domaine soumis à une norme spécifique, combiner plusieurs normes pour des domaines hybrides ou si un domaine s'appuie explicitement sur une autre norme (cas de la DO-326).
- favoriser la transposition des menaces de nature similaire entre domaines, par exemple, une attaque de mise à jour de firmwares industriels vers l'automobile avec notamment Uptane comme mesure spécialisée (Kuppusamy *et al.*, 2018).

- développer un outillage d’analyse des risques, capable de capturer les concepts essentiels en utilisant les conventions spécifiques à des domaines pour leur capture et documentation, que cela soit en termes de liste de contrôles, de niveau d’assurance ou encore de la manière de dérouler le processus d’analyse.

La démarche suivie se base sur l’identification d’un tronc commun, en premier lieu, l’ISO31000 et la norme ISO27005 plus ancienne qui a été inspirante pour les autres. Ensuite, nous procédons par enrichissement en capturant des concepts qui ont été introduits pour combler des lacunes identifiées lors de notre état de l’art. Par exemple, les notions de zones et de conduits de la norme IEC 62434 rendent possible une analyse plus modulaire en contrôlant le niveau de granularité. Ces concepts raffinent la modélisation de l’infrastructure d’EBIOS, mais avec des concepts plus forts, prenant en compte la criticité des actifs, la fonction opérationnelle, l’emplacement physique ou logique, l’accès requis ou l’organisation responsable. En termes de processus, la norme automobile ISO 21434 introduit une démarche plus fine sur les étapes d’identification des chemins d’attaque permettant de mieux quantifier le risque. Notre démarche n’est donc pas réductrice : elle vise à capturer un maximum de contributions des normes en unifiant les concepts similaires tout en prenant en compte des mécanismes et spécificités réutilisables. Ce travail ne prétend pas être parfait car il a dû réaliser certaines concessions et arbitrages.

Au niveau plus technique, le méta-modèle conceptuel est spécifié à l’aide de diagrammes de classes UML reprenant les concepts apparaissant dans les descriptions de la section 2 de l’état de l’art. Sa formalisation s’appuie aussi sur plusieurs méta-modèles déjà documentés dans des cadres spécifiques telle qu’une norme t.q. ISO 27K (Akoka *et al.*, 2018), un outil t.q. Capella (Naouar *et al.*, 2021) ou plus général comme la co-ingénierie (Bakirtzis *et al.*, 2022) ou la gestion des risques projets (Sienou *et al.*, 2009). Ces méta-modèles ne sont pas reproduits ici et notre travail se démarque (1) par son ancrage dans la cybersécurité sans être spécifique à un domaine applicatif, (2) une vision unifiant le vocabulaire des normes, (3) la prise en compte d’un contexte élargi capturant des buts de l’entreprise et les motivations de l’attaquant.

### 3.2. Méta-modèle des concepts d’analyse des risques

La figure 6 structure les biens métier et de support suivant la classification proposée par EBIOS ainsi que la structuration en zones et conduits de l’ISO 62443. La figure 7 décrit l’articulation entre les biens de l’entreprise, les propriétés de sécurité à garantir, les risques et les mesures de contrôle. Voici les caractéristiques principales :

- la partie relative à l’organisation permet la modélisation structurée des objectifs via des arbres de buts. Cela permet de relier des propriétés de sécurité aux autres buts du système et d’employer des notations et méthodes d’ingénierie des exigences orientées buts telles que KAOS (van Lamsweerde, 2009) ou i\* (Yu, Mylopoulos, 1997).
- de manière duale, la notion d’attaquant est explicitée et ses motivations sont capturées par des anti-butts appelés ici événements redoutés (terme EBIOS) qui se matérialisent via des scénarios d’attaque. Cette partie liée à la modélisation de l’attaquant

est représentée dans un ton plus foncé dans la figure 7.

- le risque lui-même est capturé selon ses dimensions d'impact (en lien avec les biens et propriétés métier) et de faisabilité (en lien avec les biens de support qui sont ceux qui peuvent contenir des vulnérabilités exploitables par des scénarios d'attaque).
- enfin, les risques de sécurité sont traités par des stratégies qui mettent en oeuvre des contrôles. Ceux-ci se situent sur différentes lignes de défense spécifiées à l'aide du framework NIST CSF (NIST, 2014).

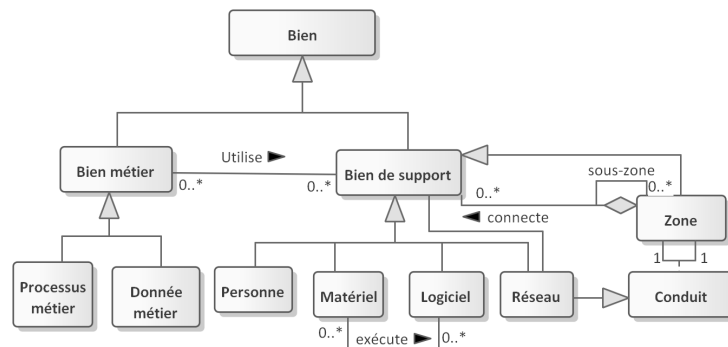


FIGURE 6. Méta-modèle de la classification des biens métier et de support

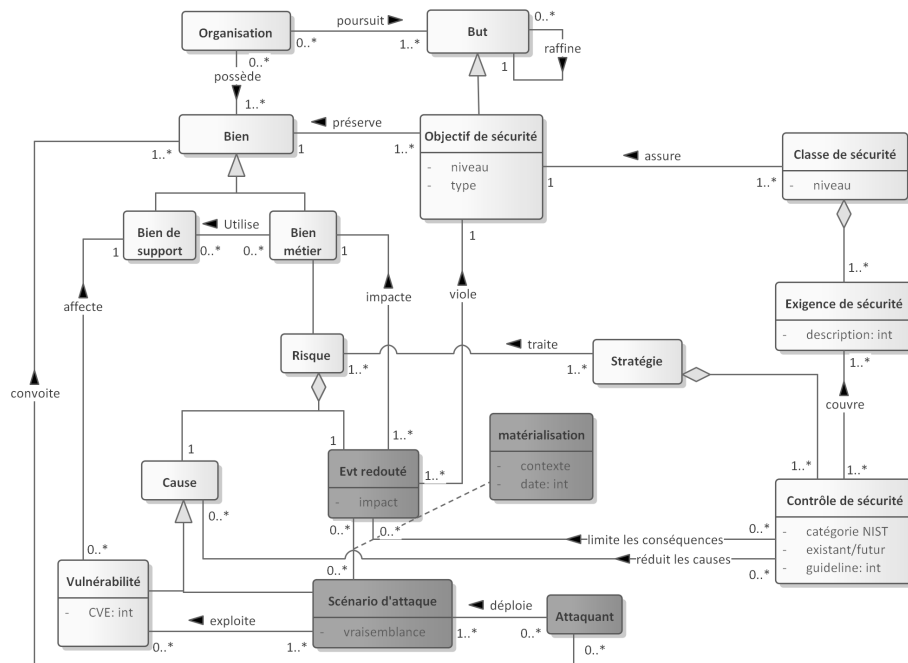


FIGURE 7. Méta-modèle des concepts de gestion des risques

### 3.3. Processus généralisé d'analyse des risques

La figure 8 représente une généralisation des processus d'analyse des risques décrits dans la section 2 à l'aide des notations BPMN. La partie gauche de la figure représente les étapes standards de l'ISO 31000.

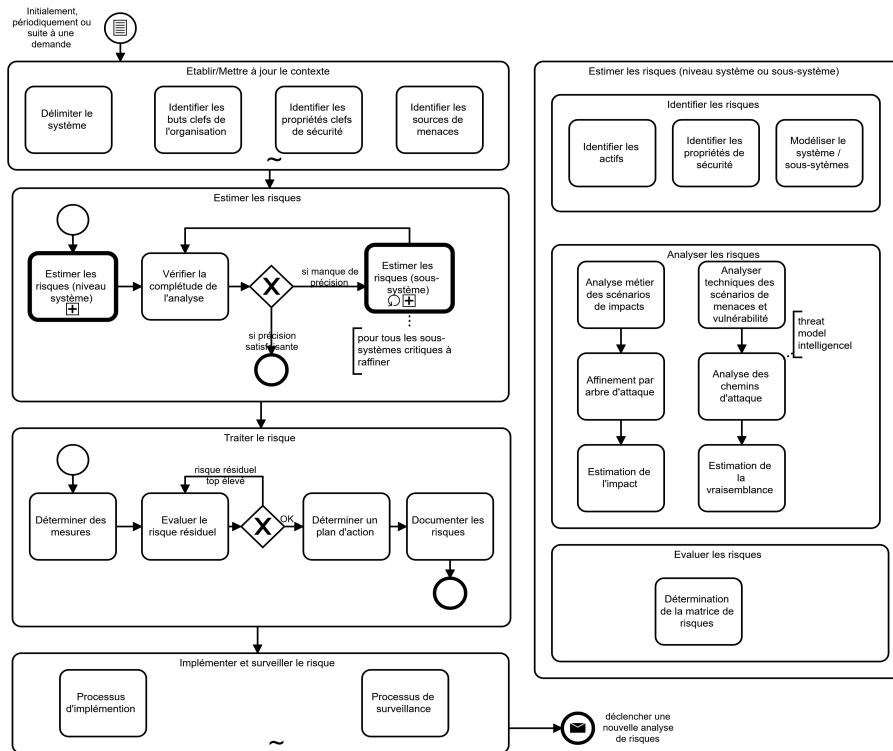


FIGURE 8. Processus d'analyse des risques

- l'étape d'établissement du contexte couvre aussi sa mise à jour lors d'une itération ultérieure. Elle comprend des activités d'identification du périmètre, des buts de l'organisation, des propriétés de sécurité et des sources de menaces.
- l'étape d'estimation des risques procède par raffinements, d'abord au niveau système, ensuite dans des sous-systèmes (ou zones) jusqu'au niveau de précision souhaité. L'étape clef (au niveau système ou sous-système) est modélisée via une « call activity » détaillée dans la partie droite de la figure 8. Elle traite séparément les aspects métier et d'infrastructure via des diagrammes spécifiques.
- l'étape de traitement des risques identifie des mesures réduisant le risque à un niveau résiduel acceptable, en plusieurs itérations si nécessaire. Les mesures sont ensuite documentées et priorisées au sein d'un planning.
- l'étape finale de mise en oeuvre et surveillance sort de la portée de l'analyse des risques mais permet d'initier une mise à jour quand c'est nécessaire.

#### 4. Discussion sur un cas d'analyse des risques dirigée par les modèles

Les méta-modèles riches proposés offrent bien sûr un cadre conceptuel clair, utile à un apprentissage de l'analyse des risques et pour faire une transposition des notions similaires entre différentes normes à faire coexister. **Cependant l'apport principal que nous illustrons ici concerne la transition vers une approche d'analyse des risques dirigée par les modèles plus précise et automatisée** que les approches manuelles basées sur des tables et documents textuels qui montrent rapidement leurs limites sur des systèmes complexes (Ponsard, Massonet, 2022). A cette fin, nous analysons un service public de traitement des eaux usées, composé d'un système de surveillance (réseau de capteurs OT) qui transmet les informations et des alarmes à une salle de contrôle informatique (IT) (qui traite les alarmes) et génère des rapports quotidiens. Atteindre les objectifs de sécurité consiste à assurer la disponibilité (éviter des pollutions et être conforme aux réglementations environnementales) et la sécurité opérationnelle (intégrité). La confidentialité n'est pas pertinente car les données sont publiques. Afin de préparer l'analyse des risques, notre diagramme de processus unifié (figure 8) indique la nécessité de modéliser le contexte, d'identifier des biens, les risques, des propriétés et de réaliser des analyses métier et techniques.

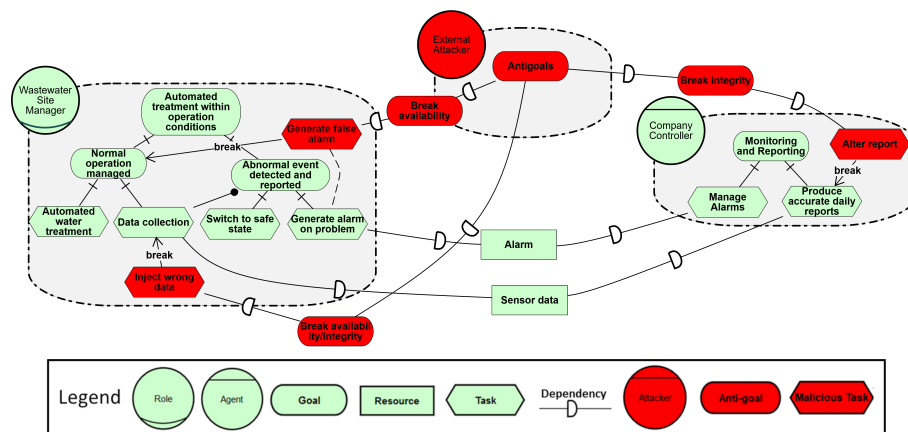


FIGURE 9. *Modèle stratégique*

**Concernant l'analyse métier pour estimer les impacts,** La figure 9 montre une modélisation dans les notations i\*. L'examen de sa légende permet immédiatement d'identifier des concepts du méta-modèle liés aux biens métier : buts, agents (zones), dépendances (conduits), attaquant, biens (ressources), événements redoutés (anti-buts) et tâches malicieuses (scénarios d'attaque). Ainsi, l'utilisation d'un outil i\* tel que piStar (Pimentel, 2018) permet immédiatement de modéliser ces concepts, les liens établis de nos modèles avec les différentes normes permettant ensuite de les transposer directement dans la norme souhaitée. Au niveau de notre modèle de processus, ceci permet de couvrir les étapes initiales liées au contexte et plus détaillées de l'analyse métier, arbre d'attaque et de l'estimation d'impact. En effet, ce modèle stratégique capture les principaux objectifs fonctionnels du système ainsi que l'intentionnalité d'un atta-

quant. Ainsi, ce raisonnement permet d'identifier les propriétés de sécurité à assurer sur des données (intégrité) et des services (disponibilité). Une analyse plus spécifique d'un sous-système (par exemple OT) peut être réalisée en fonction de sa criticité. Par exemple, un système de potabilisation soumise à la directive NIS nécessiterait une telle analyse.

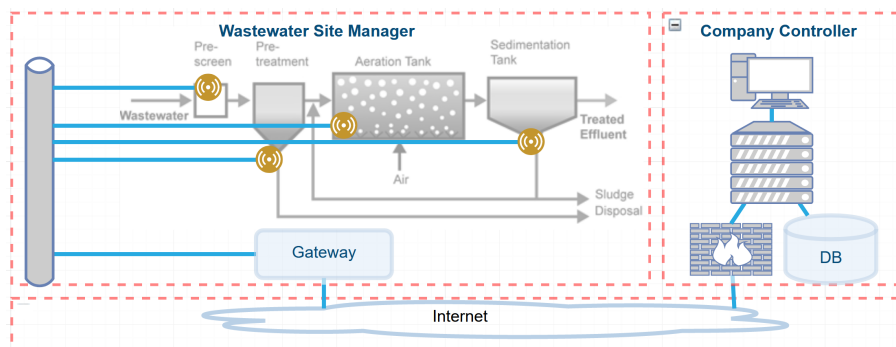


FIGURE 10. Modèle d'infrastructure réalisé avec Irius Risk

**Concernant l'analyse technique menant à l'estimation de la vraisemblance**, la figure 10 décrit l'infrastructure à l'aide de notations graphiques correspondant ici à des concepts de notre méta-modèle centré sur les biens de support. On y voit deux zones de nature différente : une zone OT avec des capteurs et un réseau IoT ainsi qu'une IT avec des serveurs et une base de données. Elles sont reliées par un conduit représenté par un réseau de communication. Cette modélisation permet des analyses automatisées, par exemple, pour identifier des vulnérabilités sur base de CVE (Common Vulnerability Enumeration) connus et des chemins d'attaques possibles, deux concepts clés du méta-modèle qui permettent d'estimer les vraisemblances des risques.

L'étape suivante du processus est d'évaluer le risque. Grâce aux concepts communs du méta-modèle, les modèles précédents peuvent être mis en correspondance notamment les acteurs du système avec les zones de l'infrastructure. Les informations respectives d'impact et de vraisemblance peuvent alors être combinées pour estimer les risques en prenant en compte des menaces multiples, la présence de mesures, etc. En manipulant les représentations des modèles (JSON, XML, EMF...) et en les transformant, il est possible d'automatiser la production de la matrice (figure 3 selon EBIOS) (Ponsard *et al.*, 2021). De cette analyse découle une série de contre-mesures. Des rapports dans des formats spécifiques à la norme cible peuvent aussi être automatisés.

Notre travail s'apparente à certains travaux, notamment UML a été étendu pour capturer les propriétés de sécurité, par exemple UMLSec (Jürjens, 2002). Les méthodes d'ingénierie des exigences dirigées par les objectifs déjà citées ont aussi été appliquées plus spécifiquement. *i\** dispose d'extensions de sécurité pour les analyses de vulnérabilité (Elahi *et al.*, 2010), Secure Tropos permet de traiter les systèmes socio-techniques (Paja *et al.*, 2013) et une ontologie spécifique à la sécurité a aussi été proposée (Sales *et al.*, 2018). Notre méta-modèle s'appuie sur une série de méta-

modèles et d'ontologies déjà publiés mais pourrait être consolidé sur base d'une revue plus systématique et approfondie de tels travaux.

## 5. Conclusion et perspectives

Dans cet article, nous avons proposé un méta-modèle de concepts et de processus donnant une synthèse unifiée de la démarche d'analyse des risques en passant en revue différentes normes de cybersécurité. Nous avons illustré les bénéfices liés à une modélisation plus précise et automatisée des risques à l'aide d'un système industriel. Nos travaux sont alignés avec d'autres propositions et permettent de combiner diverses techniques et outils, en s'affranchissant du cadre spécifique imposé par une norme.

Nos travaux futurs porteront sur l'affinement de notre méta-modèle et la mise en place d'un outillage permettant d'intégrer plus facilement divers formalismes dans notre référentiel pour y appliquer des analyses de risques. Cette approche s'inscrit aussi dans une intégration au sein d'une démarche de type DevSecOps alimentée en amont par des bases de connaissances issues de la surveillance opérationnelle, en élaborant plus cette partie de notre modèle générique de processus. En aval, les modèles produits peuvent être également exploités plus systématiquement pour alimenter des démarches de conception, de développement et de tests axés sur la sécurité.

## Remerciements

Ces travaux ont été en partie financés par les projets CYRUS (8227) et CyberExcellence (2110186). Nous remercions les partenaires industriels pour avoir partagé leurs expériences en matière d'analyse des risques.

## Bibliographie

- Akoka J., Laoufi N., Lammari N. (2018, mai). Méta modèle de la sécurité des systèmes d'information : enrichissement par le contexte. In *INFORSID 2018*. Nantes, France.
- ANSSI. (2010). *Expression des Besoins et Identification des Objectifs de Sécurité*. <https://www.ssi.gouv.fr/uploads/2011/10/EBIOS-1-GuideMethodologique-2010-01-25.pdf>.
- Bakirtzis G. et al. (2022). *An ontological metamodel for cyber-physical system safety, security, and resilience coengineering*. *Softw. Syst. Model.*, vol. 21, n° 1, p. 113–137.
- BSI. (2020). ICS Cybersecurity Assessment Framework - Suitable standards supporting a hybrid approach to risk management. *White paper*.
- David A. (2019). An Introduction to DO-326A/ED-202A – Aviation Cyber-Security Set for Engineers and Managers. <https://afuzion.com/do-326a-ed-202a-aviation-cyber-security>.
- Elahi G., Yu E., Zannone N. (2010, mars). *A vulnerability-centric requirements engineering framework*. *Requir. Eng.*, vol. 15, n° 1.
- ENISA. (2020). Threat Landscape 2020 - List of top 15 threats .

- EU. (2016). Dir. 1148 concerning measures for a high common level of security of network and information systems across the Union. <http://data.europa.eu/eli/dir/2016/1148/oj>.
- European Commission. (2016). Regulation (EU) 2016/679 - General Data Protection Regulation (GDPR). <http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679>.
- IEC. (2010a). 61508 - functional safety of electrical/electronic/programmable electronic safety-related systems. <http://www.iec.ch/functionalsafety>.
- IEC. (2010b). IEC 61508 - Functional safety of electrical/electronic/programmable electronic safety-related systems. <http://www.iec.ch/functionalsafety>.
- ISO. (2009). Risk management – vocabulary. *ISO Guide 73*.
- ISO. (2013). ISO/IEC 27000 Family - Information Security Management Systems. <https://www.iso.org/isoiec-27001-information-security.html>.
- ISO. (2018). ISO 31000, Risk management - Guidelines, provides principles, framework. <https://www.iso.org/iso-31000-risk-management.html>.
- ISO. (2020). ISO/SAE FDIS 21434 Road vehicles — Cybersecurity engineering (draft). <https://www.iso.org/standard/70918.html>.
- Jürjens J. (2002). *UMLsec: Extending UML for Secure Systems Development*. In *Uml - the unified modeling language*.
- Kuppusamy T. K., DeLong L. A., Cappos J. (2018). *Uptane: Security and customizability of software updates for vehicles*. *IEEE Vehicular Technology Magazine*, vol. 13, n° 1.
- Naouar D. et al. (2021). Towards the integration of cybersecurity risk assessment into model-based requirements engineering. In *29th int. requirements engineering conference*.
- NIST. (2014). *Cybersecurity Framework*. <https://www.nist.gov/cyberframework>.
- Paja E. et al. (2013). *Sts-tool: Specifying and reasoning over socio-technical security requirements*. In *Proc. of the 6th international i\* workshop 2013, valencia, spain, june 17-18*.
- Pimentel J. (2018). *pistar tool for i\* 2.0*. <https://www.cin.ufpe.br/~jhcp/pistar>.
- Ponsard C., Massonet P. (2022). Survey and Guidelines about Learning Cyber Security Risk Assessment. *8th Int. Conf. on Information Systems Security and Privacy, online, Feb 9-11*.
- Ponsard C., Ramon V., Touzani M. (2021). Improving Cyber Security Risk Assessment by Combined Use of i\* and Infrastructure Models. *Proc. of the 14th Int. iStar Workshop*.
- SAE. (2016). *Cybersecurity Guidebook for Cyber-Physical Vehicle Systems - J3061\_201601*. [https://www.sae.org/standards/content/j3061\\_201601](https://www.sae.org/standards/content/j3061_201601).
- Sales T. P. et al. (2018). The common ontology of value and risk. In *Conceptual modeling - 37th int. conf., ER 2018, xi'an, china, october 22-25*.
- Sienou A., Lamine E., Pingaud H. (2009, 08). A method for integrated management of process-risk. , vol. 339.
- van Lamsweerde A. (2009). *Requirements engineering - from system goals to UML models to software specifications*. Wiley.
- Yu E., Mylopoulos J. (1997, avril). Enterprise modelling for business redesign: The i\* framework. *SIGGROUP Bull.*, vol. 18, n° 1.



---

# **Pirate ta fac !**

## **Ludification de séances de cours sur la sécurité des systèmes d'information**

**Pierre-Emmanuel Arduin<sup>1</sup>, Benjamin Costé<sup>2</sup>**

1. Université Paris-Dauphine, PSL, DRM UMR CNRS 7088  
Place du Maréchal de Lattre de Tassigny, 75775 Paris Cedex 16, France  
[pierre-emmanuel.arduin@dauphine.psl.eu](mailto:pierre-emmanuel.arduin@dauphine.psl.eu)
2. Airbus Cybersecurity, Saint-Jacques-de-la-Lande, France  
[benjamin.b.coste@airbus.com](mailto:benjamin.b.coste@airbus.com)

---

*RÉSUMÉ. La sécurité des systèmes d'information s'enseigne à l'instar de toute autre activité, c'est-à-dire au travers d'exercices pratiques tels que de l'analyse de logiciels malveillants, de la détection de hameçonnage, des défis de « capture du drapeau », etc. Ces activités transforment les étudiants en apprenants actifs et facilitent leur appropriation des concepts théoriques. Nous présentons dans cet article une approche d'enseignement originale induisant un engagement accru des étudiants à qui il a été demandé de pirater des ordinateurs et de manipuler des membres de l'université. Cette approche a notamment mené à une prise de conscience sur l'importance des menaces intérieures à la sécurité des systèmes d'information, mais a aussi et surtout maintenu l'enthousiasme et l'intérêt des étudiants.*

*ABSTRACT. Teaching cybersecurity is like any other teaching activity: it requires practical exercises such as malware analysis, phishing detection, "capture the flag" challenges, etc. These activities give students the opportunity to become active learners by testing theoretic concepts and applying them in a practical way. In this article, we present an original teaching approach inducing an increased engagement from students who were asked to hack devices and deceive people within the university. Such an approach has notably led to raise awareness on the importance of insider threats in cybersecurity, but also and above all has maintained the enthusiasm and interest of students.*

*MOTS-CLÉS : Cybersécurité, Capture du drapeau, Ludification, Menaces intérieures, Influence sociale, Motivation pour apprendre.*

*KEYWORDS: Cybersecurity, Capture the flag, Gamification, Insider threats, Social influence, Learning motivation.*

---

## 1. Introduction

La sécurité des systèmes d'information peut être appréhendée d'un point de vue technique et focalisée sur les menaces externes afin de prévenir les intrusions (Hansen *et al.*, 2007) ou détecter les attaques par déni de service (Zhi-Jun *et al.*, 2012). La littérature académique aussi bien que les professionnels observent qu'une menace majeure n'est ni technique ni externe, mais provient des employés à l'intérieur même des organisations (Hassandoust *et al.*, 2020; Willison, Warkentin, 2013). Cette menace intérieure peut être intentionnelle ou non, malveillante ou non (Arduin, 2018; Leach, 2003; Loch *et al.*, 1992; Warkentin, Willison, 2009), ce qui a conduit les professionnels, chercheurs et enseignants de l'enseignement supérieur à considérer la cybersécurité ou sécurité des systèmes d'information comme un phénomène social (McAlaney, Benson, 2020).

L'enseignement de la cybersécurité ne se réduit pas à l'enseignement de techniques de chiffrement/déchiffrement, d'analyse de réseau ou d'attaques par force brute. Venkatesh *et al.* (2003) ont montré que l'intention des individus de passer à l'acte est notamment déterminée par le fait qu'ils pensent que les autres individus soutiennent ou condamnent l'acte en question. Pour McAlaney, Benson (2020), cela démontre le besoin de comprendre comment non seulement les individus perçoivent les risques de cybersécurité, mais aussi comment ils pensent que les autres perçoivent ces risques. Dans cet article, nous approfondissons cette idée en présentant comment entretenir l'enthousiasme et l'investissement des étudiants lors de séances de cours sur la sécurité des systèmes d'information nécessitant une maîtrise de l'ingénierie aussi bien technique que sociale.

Dans la deuxième section de cet article, nous présentons les théories mobilisées : d'abord des concepts fondamentaux de la sécurité des systèmes d'information à partager avec les étudiants, menaces extérieures et intérieures, ensuite une revue de la littérature en pédagogie par la ludification. Dans la troisième section, nous présentons des scénarios conçus pour des cours de sécurité des systèmes d'information en deuxième année de master : d'abord une description des séances sur les menaces extérieures et intérieures, ensuite une description de l'exercice de capture du drapeau, enfin une discussion sur les limites et les implications éthiques de l'approche considérée dans ce travail. En effet, cet article vise à partager des scénarios pédagogiques ludiques pour enseigner la sécurité des systèmes d'information et stimuler la motivation des étudiants : pirate ta fac !

## 2. Concepts théoriques et revue de la littérature

Dans cette section, nous présentons d'abord les éléments cruciaux de la sécurité des systèmes d'information à partager avec les étudiants dans le cadre du cours considéré dans cet article : les menaces extérieures et intérieures à la sécurité des systèmes d'information. Ensuite, nous proposons des éléments de la littérature sur la ludification comme approche pédagogique.

### 2.1. Maîtriser les technologies, une condition préalable à la cybersécurité

Un système d'information (SI) n'est pas uniquement composé d'appareils technologiques ; il inclut également les humains (Reix, 2000) qui concourent activement à la sécurité du système global. Le SI est ainsi menacé à double titre : d'une part, les attaquants externes utilisent leurs propres ressources pour pénétrer puis compromettre le SI, d'autre part, les utilisateurs internes peuvent menacer intentionnellement ou non la sécurité du système. De plus, l'émergence de l'Internet des objets (IoT) (Atzori *et al.*, 2010), du Bring Your Own Device (BYOD) (Thomson, 2012) et des modèles de confiance zéro (Ward, Beyer, 2014) renforce l'importance de considérer la sécurité des deux points de vue.

L'essor des appareils connectés de tous types (smartphones, assistants personnels intelligents, etc.) amène de nouvelles habitudes et un besoin croissant d'accès numérique. Étant connectés numériquement à presque chaque instant, nos données sont en même temps vulnérables. Les utilisateurs peuvent consolider la protection de ces nouveaux appareils de trois manières. Premièrement, le recours à des produits payants d'éditeurs renommés tels que des antivirus, des réseaux privés virtuels (VPN), des pare-feu, des protections Web, etc. qui offrent un niveau de sécurité élevé (Ahvanooy *et al.*, 2017). Deuxièmement, l'utilisation de produits gratuits dont la contrepartie est leur faible efficacité. Il peut s'agir de versions de produits bien connus moins riches en fonctionnalités que leurs versions payantes ou de nouvelles solutions d'éditeurs à la conquête d'un marché. Cependant, il peut aussi s'agir de faux produits qui installent des *adwares* et/ou des *pop-ups* pour collecter (voire voler) des données personnelles en même temps qu'ils introduisent de nouvelles vulnérabilités (Wu *et al.*, 2014). Troisièmement, la configuration personnalisée des périphériques sans recours à un logiciel de sécurité supplémentaire. Cela nécessite cependant une sensibilisation aux problématiques de sécurité et une connaissance approfondie de leurs aspects techniques. Par exemple, la plupart des gens laisse les interfaces sans fil activées, diffusant leurs données et encourageant les attaquants opportunistes.

Chaque option comporte ses propres risques et la plupart des utilisateurs n'ont pas l'utilité de comprendre les éléments techniques sous-jacents bien qu'ils aient besoin de choisir les moyens de sécurité appropriés. En tant qu'informaticiens, les étudiants doivent comprendre quels sont les risques de leurs décisions, surtout lorsqu'ils mettent leurs compétences au service d'une entreprise dès lors que leurs décisions affectent plusieurs utilisateurs. De plus, les cybercriminels agissent différemment lorsqu'ils ciblent le SI des entreprises plutôt que les appareils personnels. Le cours que nous proposons vise à aider les étudiants à agir de manière appropriée pour surmonter les risques personnels ou à l'échelle de l'entreprise. Dans cette optique, nous proposons de nous concentrer sur ces menaces externes à travers trois grands objectifs pédagogiques expliqués ci-après :

1. Le premier objectif est de penser comme un cybercriminel afin d'identifier et anticiper les vulnérabilités et les menaces. En effet, un esprit criminel n'est pas intuitif par nature car il contourne les outils ou les habitudes pour obtenir ce qu'il recherche.

Les phreakers utilisaient un sifflet pour passer des appels téléphoniques gratuits (Mitnick, Simon, 2003), certains pirates utilisent des LED optiques pour exfiltrer les données d'ordinateurs isolés (Guri *et al.*, 2016) ou même transforment les alimentations électriques en haut-parleurs (Guri, 2020). Les appareils peuvent ainsi être détournés de leur fonction première pour réaliser des actions non-désirées par les concepteurs voire strictement proscrites par ceux-ci. Le premier objectif de notre cours est de partager une telle idée avec nos étudiants.

2. Le deuxième objectif est que les étudiants puissent comparer leur approche avec celles d'attaquants réels. Ils apprennent par ce biais ce qui est réellement utilisé par les attaquants au travers d'exemples d'attaques réussies.

3. Le troisième objectif est de reconstruire une séquence d'événements et de trouver des éléments d'information sur un attaquant au travers de leur compréhension du modus operandi criminel.

Ces trois objectifs concourent à fournir aux étudiants une juste appréciation des risques et des procédures criminelles.

## 2.2. Manager les individus, un besoin grandissant pour la cybersécurité

Les conséquences d'une attaque mises en évidence très tôt par Loch *et al.* (1992) restent les mêmes de nos jours : (1) divulgation d'informations monnayables, (2) modification ou (3) destruction d'informations sensibles et (4) déni de service, en entravant l'accès aux ressources. En particulier, ces auteurs ont souligné l'existence de menaces extérieures et intérieures pour la sécurité des SI (Fig. 1).

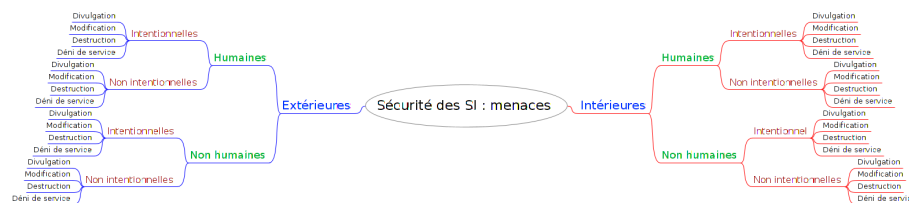


FIGURE 1. Taxonomie des menaces à la sécurité des SI (inspiré de Loch *et al.*, 1992)

Johnston, Warkentin (2010) ont noté que l'influence sociale a un impact important sur les intentions des utilisateurs finaux en matière de sécurité. Pour Venkatesh *et al.* (2003), l'intention d'avoir un comportement en particulier dépend de la perception que vous avez que les autres soutiendront ou condamneront ce comportement. Dois-je faire part de mes doutes sur ce courriel suspect à mes collègues ? Dois-je faire confiance à cet assistant qui me presse de lui envoyer un fichier ? Des auteurs tels que Arduin (2021); Tsohou *et al.* (2015) ont montré que les biais cognitifs et culturels peuvent être activés de manière malveillante pour influencer les intentions des utilisateurs du SI quant à la sécurité.

Les menaces intérieures peuvent être classées selon deux dimensions : (1) le caractère intentionnel de la menace et (2) sa malveillance (Willison, Warkentin, 2013).

Pour les employés utilisateurs d'un SI et constituant un point d'entrée dans le système, les menaces intérieures peuvent ainsi être (Arduin, 2018) :

1. *non-intentionnelles* : actions erronées d'employés inexpérimentés, négligents ou influencés ; par exemple, des clics inattentifs, des erreurs de saisie, des suppressions accidentelles de données sensibles, etc. (Stanton *et al.*, 2005),

2. *intentionnelles et non-malveillantes* : actions délibérées d'employés ayant un bénéfice mais sans volonté de nuire ; par exemple, différer les sauvegardes, choisir un mot de passe faible, laisser les portes ouvertes lors de discussions sensibles, etc. (Guo *et al.*, 2011),

3. *intentionnelles et malveillantes* : actions délibérées d'employés ayant la volonté de nuire ; par exemple, divulgation de données sensibles, introduction de logiciels malveillants, etc. (Shropshire, 2009).

Le cours que nous avons proposé intègre les aspects technologiques de la cybersécurité, tels que présentés dans la section 2.1, mais aussi les aspects sociaux et comportementaux de la cybersécurité s'appuyant sur ces trois catégories de menaces internes. En effet, la formation des individus est nécessaire pour contrer les attaques reposant sur des techniques d'ingénierie sociale et de manipulation (Campbell, 2019). Mitnick, Simon (2003) définissent l'ingénieur social comme étant un attaquant qui cible un utilisateur légitime duquel il obtient un moyen d'accès direct (droits d'accès, lien nuisible visité, etc.) ou indirect (informations vitales, relation de confiance, etc.) au système. Le contenu de la section 3 présente comment nos étudiants ont appris à penser en utilisant des techniques d'ingénierie sociale. Ils ont été impliqués dans le cours par la ludification des séances.

### **2.3. La ludification, un vecteur d'implication des étudiants dans l'enseignement**

Si l'enseignement repose sur le partage des connaissances, il se heurte parfois à certaines difficultés : dans le cadre de l'enseignement en général et de l'enseignement supérieur en particulier, le désengagement induit par les méthodes pédagogiques traditionnelles est critiqué par des auteurs comme Siala *et al.* (2019). D'autres comme Mustar (2009) considèrent que les étudiants n'y sont pas assez remis en cause et que ces méthodes ne partagent pas les connaissances pratiques.

« J'entends et j'oublie, je vois et je me souviens, je fais et je comprends » : cette phrase, attribuée à Confucius au II<sup>e</sup> siècle avant J.C., illustre bien l'idée du « *learning by doing* » formalisée par Kolb (1984). Cette idée est plutôt éloignée de l'apprentissage par frustration cognitive proposée par Cangelosi, Usrey (1970). Apprendre en agissant conduit non seulement à améliorer la pensée critique et les compétences de résolution pratiques (Kapp, 2012), mais aussi – et peut-être le plus important – la connaissance apprise devient « plus intéressante » (Tobias *et al.*, 2014).

Des méthodes d'enseignement traditionnelles aux cours en ligne ouverts et massifs (MOOC), l'éventail des façons de partager les connaissances et d'enseigner est assez large. Si les MOOCs fournissent un contenu pédagogique en ligne et sont pré-

sentés comme permettant aux individus d'apprendre de manière autonome chez eux (Razmerita *et al.*, 2019), il convient de rappeler que l'intention de rester dans le cours doit être mieux monitorée et analysée, ce qui a été particulièrement vrai pendant les confinements successifs dus au COVID-19 (Prenkaj *et al.*, 2020).

Snyder (2018) a observé des expériences d'enseignement de la cybersécurité avec des *escape games*, des jeux d'évasion. Il a humblement conclu que ces expériences peuvent fonctionner – ou pas – mais que dans tous les cas les participants ont passé un bon moment. Bruguier *et al.* (2020) a identifié trois composants importants lors de la conception d'un jeu d'évasion pour l'enseignement de la sécurité matérielle : (1) l'importance du scénario, (2) la posture de l'enseignant et (3) le besoin d'une *debriefing*, une réunion-bilan. Silic, Lowry (2020, p. 131) vont plus loin lorsqu'ils proposent de définir la *gamification*, la ludification, de la sécurité comme un moyen de motiver les employés afin d'encourager l'apprentissage, l'efficacité et la conformité avec les initiatives de sécurité en utilisant des artefacts et des processus inspirés par le jeu.

La ludification de séances de cours sur la sécurité des systèmes d'information apparaît alors plutôt évidente face à la demande de nouvelles méthodes facilitant l'appréciation collective des objectifs de sécurité. Même si une telle question a déjà été partiellement abordée à la fin des années 1990 avec la montée des *hackathons*, des événements intensifs de programmation informatique (Maaravi, 2020). Une étude réalisée par Briscoe, Mulligan (2014) conclut que les 150 répondants participent à des *hackathons* pour l'apprentissage (86%), le réseautage (82%), ou pour faire avancer le changement social (38%). En effet, ces auteurs rappellent que l'une des origines des *hackathons* est de hacker du code dans un but d'amélioration sociale.

### 3. Proposition : un cours « ludique » sur la sécurité des systèmes d'information

Plutôt que de former des experts en cybersécurité, le cours que nous proposons vise à assurer une compréhension fondamentale des concepts de sécurité des systèmes d'information les plus courants tout au long du cycle de vie de la sécurité : identification, protection, détection, réponse et récupération (NIST, 2018). Pour chaque partie, une brève description des problèmes rencontrés par les experts en cybersécurité est présentée, en s'appuyant aussi bien sur la littérature en la matière que sur nos propres expériences. Les bases de la cryptographie font également partie du cours. Chaque concept présenté est abordé aussi bien du côté attaquant (« équipe rouge ») que défenseur (« équipe bleue ») au regard des objectifs pédagogiques 1 et 3 (*cf.* section 2.1) ainsi que pour des raisons éthiques (Mirkovic, Peterson, 2014).

Le cours est organisé en trois parties successives. Trois cours magistraux pour commencer, sur les systèmes d'information, l'ingénierie sociale, les menaces intérieures, le cycle de vie de la sécurité et la cryptographie. Ces séances au contenu classique ne sont pas présentées dans cet article. S'ensuivent trois Travaux Dirigés (TD) répartis en deux thèmes : menaces extérieures et intérieures. Les étudiants choisissent eux-mêmes le thème qu'ils souhaitent suivre mais la répartition observée ces dernières années est plutôt équilibrée. Le cours se termine par un exercice de *Capture*

*The Flag* (CTF), capture du drapeau, introduit après les cours théoriques qui se tient en parallèle des séances de TD pour tous les étudiants et dont le contenu est détaillé dans la section 3.3.

### 3.1. De la maîtrise des technologies aux menaces extérieures

Après les cours théoriques, les étudiants sont répartis en groupes pour des cours dirigés. Chaque leçon pratique repose sur un scénario où les étudiants se voient confier un rôle de sécurité différent par rapport aux objectifs pédagogiques : tutoriel 1 – cybercriminel, tutoriel 2 – analyste des cybermenaces et tutoriel 3 – gestionnaire d’incidents. À la fin de chaque leçon, les étudiants doivent rédiger un rapport décrivant leur méthodologie. Les étudiants étant supposés novices en sécurité, leur sensibilisation aux problèmes de sécurité est privilégiée par rapport aux compétences techniques. Cela a une influence sur la conception des leçons : le premier scénario ne nécessite qu’une enquête sur le Web, le second nécessite une compréhension du vocabulaire techniques et le troisième nécessite des compétences techniques de base.

#### 3.1.1. TD 1 – Hackons-le !

Dans le premier scénario, les élèves doivent imaginer un chemin pour compromettre un réseau commun, représenté sur la figure 2, comportant notamment une zone démilitarisée (DMZ) exposée sur Internet ainsi que différentes zones internes. Ils sont invités à s’inspirer du framework MITRE ATT&CK (Strom *et al.*, 2018) décrivant la méthodologie d’un attaquant. Ce framework est notamment utilisé en cas de détection d’attaque (Al-Shaer *et al.*, 2020) pour identifier les avancées des attaquants. Les exemples présentés sur le site du MITRE (cf <https://attack.mitre.org/matrices/enterprise/>) amènent les étudiants à découvrir plusieurs méthodologies d’attaquants.

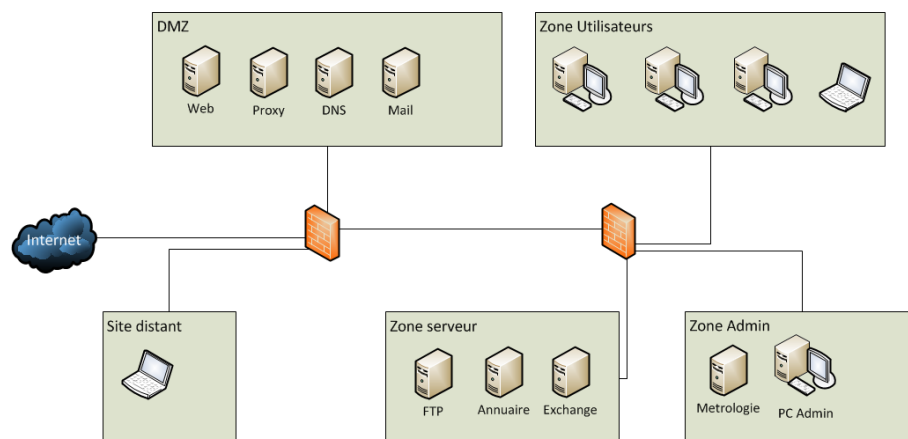


FIGURE 2. SI ciblé par les étudiants lors d’un TD sur les menaces extérieures

### 3.1.2. TD 2 – Connaître les attaquants

Dans le deuxième scénario, les étudiants se voient confier le rôle d'experts en Cyber Threat Intelligence (CTI). Chaque groupe doit analyser un rapport de menace différent, résumer la méthodologie employée par les attaquants et proposer des contre-mesures. Les rapports de menace utilisés en support sont de véritables rapports portant sur des attaques réelles. Si ce TD est principalement documentaire, il est intéressant de noter qu'il mobilise de nombreuses notions techniques, parfois avancées, qui poussent les étudiants à confronter leurs connaissances à la réalité, notamment en reconsidérant l'impact de la présence d'un antivirus (souvent contourné) ou de celle des mises à jour de sécurité (pas toujours appliquées).

### 3.1.3. TD 3 – Trouvez le chat

Le troisième scénario est beaucoup plus technique. Les élèves doivent récupérer des éléments d'information sur un kidnappeur à l'aide de compétences en criminalistique numérique et en réponse aux incidents. Dans ce scénario, le chat de leur manager a été kidnappé (voir Fig. 3), et ils doivent trouver l'emplacement du kidnappeur et recouvrer les données secrètes chiffrées. Les étudiants reçoivent trois fichiers : des données secrètes chiffrées et deux demandes de rançon pour le chat et les données.

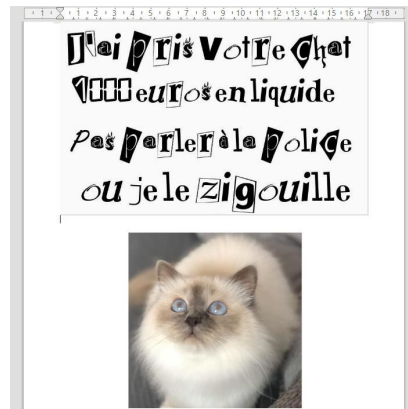


FIGURE 3. TD sur les menaces extérieures : demande de rançon avec données cachées.

Chaque fichier représente un défi différent : (i) le fichier de la rançon du chat contient l'emplacement probable du kidnappeur, caché dans les métadonnées ; (ii) Le deuxième fichier de rançon contient le mot de passe sécurisé qui est nécessaire pour déchiffrer les données secrètes (alias le troisième fichier) et qui ne peut être deviné ni forcé.



### 3.2. Du management des individus aux menaces intérieures

Les étudiants ayant choisi de rejoindre les TDs sur les menaces intérieures sont également répartis en groupes de quatre ou cinq personnes. Chaque séance aborde un cas spécifique de menace intérieure comme mentionné dans la section 2.2 : TD 1 – le non-intentionnel, TD 2 – l'intentionnel et non-malveillant et TD 3 – l'intentionnel et malveillant. À la fin de chaque séance, les étudiants rédigent un compte-rendu qui est évalué par l'enseignant et discuté en début de séance suivante.

#### 3.2.1. TD 1 – Les menaces intérieures non-intentionnelles

Le premier TD est organisé en quatre temps : (i) les étudiants sont confrontés individuellement à des courriels de hameçonnage dans une salle d'expérimentation avec des murs de protection anti-copie (voir fig. 4.a), ils ne disposent d'aucune information sur la véracité des courriels et doivent indiquer les zones et éléments les amenant à faire confiance ou à se méfier ; (ii) les résultats de l'expérience sont discutés avec l'enseignant, en particulier les zones identifiées comme inspirant le plus la confiance et la méfiance (voir fig. 4.b) ; (iii) les étudiants passent ensuite en mode projet et doivent concevoir un courriel de hameçonnage « optimisé » en s'appuyant sur les résultats de l'expérience et expliquer leurs choix dans un compte-rendu écrit et noté (des exemples sont disponibles dans Arduin 2021) ; (iv) les travaux sont enfin présentés et discutés.



FIGURE 4. TD sur les menaces intérieures – Hameçonne-moi si tu peux! Murs anti-copie (a) et courriel de hameçonnage avec zones de confiance (b). Les zones les plus chaudes sont les plus sélectionnées comme inspirant la confiance.

#### 3.2.2. TD 2 – Les menaces intérieures intentionnelles et non-malveillantes

Pendant le deuxième TD, la classe est divisée en deux groupes : (A) le groupe des Responsables de la Sécurité des Systèmes d'Information (RSSIs) paranoïaques et (B) le groupe des employés malins et paresseux, spécialistes des solutions de contournement et du moindre effort. Comme le lecteur l'aura peut-être compris, le groupe (A) représente les RSSIs déployant des Politique de Sécurité du Système d'Information (PSSIs) très contraignantes, tandis que le groupe (B) représente les employés contournant ces PSSIs et créant des menaces intérieures intentionnelles et non-malveillantes.

Par exemple : imposer de changer son mot de passe à des fréquences farfelues peut conduire à les écrire sur des post-its, ou encore concevoir des procédures de sauvegarde complexes car trop sécurisées peut conduire à différer les sauvegardes, etc. Ce deuxième TD est organisé en trois temps : (i) une phase au cours de laquelle les étudiants du groupe (A) préparent des PSSI complexes, très contraignantes, voire sciemment farfelues, alors que les étudiants du groupe (B) anticipent ces PSSI singulières et comment ils pourraient les contourner, un compte-rendu qui sera noté est demandé avant de passer au temps suivant ; (ii) une phase de *battle*, confrontation au cours de laquelle des étudiants du groupe (A) vont avancer leurs PSSI et des étudiants du groupe (B) exposer leurs solutions de contournement ; (iii) une phase de *débriefing*, discussion-bilan où les PSSI et les solutions de contournement sont discutées.

### 3.2.3. TD 3 – Les menaces intérieures intentionnelles et malveillantes

Le troisième et dernier TD se concentre sur l'une des plus grandes craintes des RSSIs : les employés ayant des accès privilégiés qui deviennent des menaces intérieures intentionnelles et malveillantes. Durant ce TD, les groupes d'étudiants travaillent selon une procédure utilisée habituellement en *design thinking*, conception créative : la méthode DKCP (*Define, Knowledge, Concept, Project*) (Damart et al., 2018). Les étudiants imaginent et conçoivent des pratiques parfois connues parfois inconnues où les utilisateurs deviennent intentionnellement des attaquants ayant la volonté de nuire. Après la phase individuelle de définition et de recueil de connaissances (D, K), les étudiants se mettent en groupe et échangent leurs idées pour proposer de nouveaux concepts (C) de menaces intérieures intentionnelles et malveillantes ; ils décrivent enfin la procédure comme s'il s'agissait d'un projet (P). À la fin du TD un compte-rendu est demandé, puis ces procédures sont présentées aussi bien que le processus de conception qui y a conduit. Une discussion-bilan vient conclure la séance, pendant laquelle les groupes sont invités à réfléchir aux contre-mesures possibles aux menaces identifiées.

### 3.3. L'exercice de capture du drapeau : accéder au sujet d'examen ?

La dernière partie de ce cours est conçue comme un défi de capture du drapeau (Snyder, 2018, voir section 2.3). Les étudiants doivent retrouver une partie de l'examen final cachée dans les murs de l'Université. Le défi est divisé en plusieurs parties, certaines étant communes à tous les groupes et certaines dépendant du groupe. Une cinquantaine d'étudiants en deuxième année de master MIAGE suivent ce cours chaque année. Ils ont été affectés au hasard à un groupe et plusieurs parties de l'examen final, les drapeaux, ont été cachées dans l'Université. Ainsi, tous les groupes ne cherchent pas un drapeau unique, mais deux ou trois groupes sont tout de même en concurrence pour un même drapeau. Nous avons observé que certains groupes ont capturé plus d'un drapeau.

### 3.3.1. Partie 1 – Pirater un ordinateur portable (commun à tous les groupes)

Dans la première partie, un ordinateur portable est remis aux étudiants. Plusieurs vieux postes ont été récupérés au service informatique de l'Université et préparés en amont. Ceux-ci sont verrouillés et les étudiants n'ont aucun indice supplémentaire ni soutien de la part des enseignants. Tout comme une attaque réelle, ils doivent déverrouiller la session utilisateur afin de trouver des indices sur le lieu dans lequel se trouve leur drapeau cible, une partie du sujet d'examen. Rappelons que Bruguier *et al.* (2020, voir section 2.3) a mis en évidence trois composantes importantes de l'apprentissage par le jeu : (1) l'importance du scénario, (2) la posture de l'enseignant, et (3) le besoin d'un *débriefing*, une réunion-bilan. Comme le lecteur peut le deviner, l'importance du scénario était ici cruciale, tout comme la posture que nous avons adoptée lorsque nous avons demandé aux étudiants de pirater ces postes. Nous avons été très clairs depuis le début avec eux en leur expliquant que l'examen final sera très, très difficile et qu'ils pourraient en retrouver des parties que nous avions cachées à l'Université en découvrant des indices après avoir déverrouillé l'ordinateur portable. Bien sûr, ils ont été évalués sur la méthode qu'ils ont employée pour le faire. Le temps d'accès à l'ordinateur portable était limité par groupe et surveillé. Enfin, les groupes ont dû préparer un compte-rendu sur les vulnérabilités qu'ils ont exploitées, qu'elles soient extérieures ou intérieures. Il est important de souligner que bien que les étudiants soient totalement libres sur les méthodes à employer, il leur est rappelé que toute action entreprise dans le cadre de ce CTF relève de leur seule responsabilité.

Plusieurs indices ont été laissés comme un numéro de compte Twitter caché dans un faux code-barres que nous avons préparé (fig. 5.a) ou un indice donné par l'écran de verrouillage de Microsoft Windows après des essais infructueux (fig. 5.b). Une fois la session déverrouillée, une analyse stéganographique des fichiers des utilisateurs locaux était nécessaire pour révéler des adresses URL indiquant la cible suivante aux étudiants en fonction de leur numéro de groupe. Commence alors la deuxième phase : Le jeu virtuel de cette première partie de l'exercice de capture du drapeau, centré sur les aspects technologiques de la cybersécurité (voir section 2.1), devient un jeu réel, centré sur les aspects managériaux et comportementaux de la cybersécurité (voir section 2.2).

### 3.3.2. Partie 2 – Manipuler les individus (différent en fonction des groupes)

Les adresses URL découvertes dans la première partie donnent accès à différents défis selon les groupes et amènent les étudiants à rechercher des parties de l'examen final dans l'Université. Pour certains groupes, du code HTML caché contenait des informations sur une personne à retrouver. Pour d'autres, une page Web demandait un « identifiant administrateur » relativement trivial et facile à deviner. Certains groupes ont dû ouvrir des bureaux verrouillés avec des codes ou des clés (fig. 5.c). D'autres ont dû découvrir des dates de naissance de personnels de l'Université pour pouvoir accéder à des salles de conférence. Ensuite, une feuille de papier était cachée dans le bureau ou la salle en question (derrière la porte, derrière une affiche de consignes incendie, sur le bureau, etc.). Les étudiants ont eu à exploiter des menaces intérieures

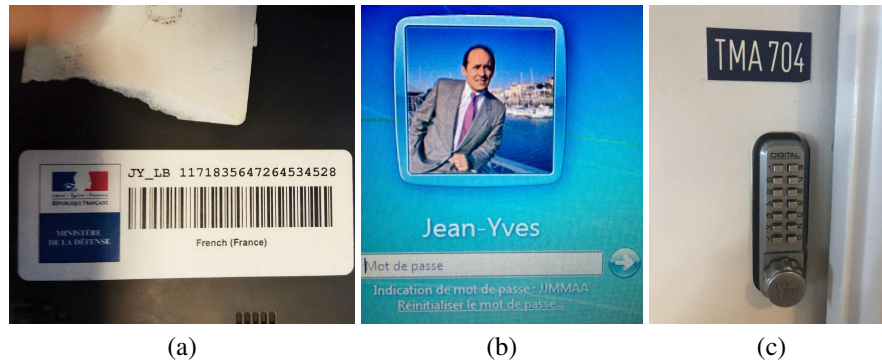


FIGURE 5. Faux code-barres (a) et écran verrouillé (b) : pirater un poste.  
Serrure de porte avec code (c) : manipuler des employés.

(voir section 2.2) en incitant le personnel de l'Université à leur laisser accéder à des zones sécurisées. Là encore, le temps était limité depuis la fin de la première partie et ils ont dû préparer un compte-rendu noté sur les menaces qu'ils ont exploitées.

### 3.4. Discussion, limites et implications éthiques

L'approche pédagogique proposée ici repose évidemment sur du « *learning by doing* » (Kolb, 1984, voir section 2.3). Des auteurs comme Sagarin, Mitnick (2012) appelaient déjà il y a dix ans à une meilleure formation des internautes sur les techniques de manipulation employées par les attaquants. Pour d'autres comme Fiske, Taylor (2013), les individus ont tout simplement du mal à se lancer dans une réflexion approfondie et fatigante, ce qui les rend plus vulnérables aux armes d'influence en ligne (Muscanell *et al.*, 2014) et rend nécessaire la formation à ce type de menace.

La section 2.1 a mis en avant le lien de plus en plus fort entre la sécurisation personnelle et son impact sur la sécurisation d'un périmètre plus large (e.g. réseau entreprise). Le premier volet, généralement connu par les étudiants, est confronté tout au long du cours aux méthodes criminelles (objectifs du cours décrits en fin de section). Par exemple, le TD1 (cf. 3.1.1) aborde la méthodologie des attaquants et pousse les étudiants à se placer d'un point de vue offensif. Le TD débute par une explication de l'architecture de l'infrastructure technique ciblée, notamment la criticité des différentes zones présentées. Il est assez facilement admis par les étudiants que la distance à Internet est révélatrice de l'importance des différents équipements, ce qui les amène rapidement à considérer le (*spear*) *phishing* comme premier moyen d'intrusion, lequel est effectivement le vecteur initial d'infection le plus couramment observé dans le contexte d'attaques ciblées. Un suivi à plus long terme des étudiants permettrait de mieux comprendre quelles connaissances ils ont retenu et sont en capacité de mobiliser dans leurs contextes professionnels spécifiques.

Nous avons observé avec des questionnaires de satisfaction en fin de cours que les étudiants étaient significativement satisfaits. Ils ont apprécié le challenge (« c'était particulièrement intéressant de trouver les indices tout seul ») et la formation technique (« peut-être approfondir un peu ce qui se passe sur le réseau, les ports, etc. »). L'ensemble des étudiants a apprécié le contrôle continu (« c'est une méthode d'évaluation qui me convient ») et nous avons même observé des étudiants partager leurs réponses ou utiliser des techniques d'ingénierie sociale sur d'autres étudiants ou du personnel administratif de l'Université, préalablement informé bien sûr mais sans détails sur l'exercice pour ne pas biaiser leur comportement.

Le risque de double usage de cette approche pédagogique ne peut pas être négligé (Rath *et al.*, 2014). Même s'il existe un risque réel que des attaquants utilisent le matériau présenté dans cet article comme un guide pratique, nous considérons que les avantages en termes de formation l'emportent sur les risques. En effet, nous soutenons que la formation des étudiants, des employés et des citoyens reste l'une des défenses les plus efficaces. Cette idée est d'ailleurs renforcée par l'explosion des cyberattaques utilisant l'ingénierie sociale du fait de l'épidémie de COVID-19 (Lallie *et al.*, 2020).

#### 4. Conclusion

Dans cet article, nous avons proposé une approche pédagogique ludique pour des cours de sécurité des systèmes d'information renforçant l'idée que l'enseignement supérieur pouvait être le lieu d'expérimentations pédagogiques fertiles pour susciter et entretenir un engagement accru des étudiants : pirate ta fac !

Dans la deuxième section, nous avons présenté les menaces extérieures et intérieures à la sécurité des systèmes d'information, ainsi que des initiatives pédagogiques innovantes telles que la ludification. Dans la troisième section, nous avons présenté la structure du cours proposé, des travaux dirigés et l'exercice de capture du drapeau. Une discussion sur les limites et les implications éthiques de la pédagogie proposée est venue conclure le propos.

Pour des auteurs tels que Shah *et al.* (2019, p. 1128), la communauté universitaire doit s'efforcer de travailler avec toutes les parties pour offrir les meilleures pratiques. En effet, la nécessité de sensibiliser et de former les étudiants aux menaces de sécurité des systèmes d'information est plus que jamais cruciale : les individus sont connectés en permanence au système et constituent des points d'entrée sensibles aux armes d'influence en ligne (Muscanell *et al.*, 2014), ce à quoi il convient de les éduquer au plus tôt.

Si l'objet de ces travaux est avant tout de partager une initiative pédagogique innovante, il reste à mieux apprécier, dans des travaux futurs, son efficacité du point de vue de l'apprentissage.

### Bibliographie

- Ahvanooy M. T., Li Q., Rabbani M., Rajput A. R. (2017). A survey on smartphones security: Software vulnerabilities, malware, and attacks. *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 8, n° 10, p. 30–45.
- Al-Shaer R., Spring J. M., Christou E. (2020). Learning the associations of MITRE ATT&CK adversarial techniques. *ArXiv*.
- Arduin P.-E. (2018). *La menace intérieure*. ISTE Éditions.
- Arduin P.-E. (2021). A cognitive approach to the decision to trust or distrust phishing emails. *International Transactions in Operational Research*, vol. à paraître.
- Atzori L., Iera A., Morabito G. (2010, oct). The internet of things: A survey. *Computer Networks*, vol. 54, n° 15, p. 2787–2805.
- Briscoe G., Mulligan C. (2014). Digital innovation: The hackathon phenomenon.
- Bruguier F., Lecointre E., Pradarelli B., Dalmaso L., Benoit P., Torres L. (2020). Teaching hardware security: Earnings of an introduction proposed as an escape game. In *International conference on remote engineering and virtual instrumentation*, p. 729–741.
- Campbell C. C. (2019). Solutions for counteracting human deception in social engineering attacks. *Information Technology & People*.
- Cangelosi V. E., Usrey G. L. (1970). Cognitive frustration and learning. *Decision Sciences*, vol. 1, n° 3-4, p. 275–295.
- Damart S., David A., Klasing Chen M., Laousse D. (2018, juin). Turning managers into management designers: an experiment. In *XXVIIème conférence de l'AIMS*. Montpellier, France.
- Fiske S. T., Taylor S. E. (2013). *Social cognition: From brains to culture*. Sage.
- Guo K., Yuan Y., Archer N., Connely C. (2011). Understanding nonmalicious security violations in the workplace: a composite behavior model. *Journal of Management Information Systems*, vol. 28, n° 2, p. 203-236.
- Guri M. (2020). Power-supplay: Leaking data from air-gapped systems by turning the power-supplies into speakers. *arXiv preprint arXiv:2005.00395*.
- Guri M., Hasson O., Kedma G., Elovici Y. (2016, dec). An optical covert-channel to leak data through an air-gap. In *2016 14th annual conference on privacy, security and trust (PST)*. IEEE.
- Hansen J. V., Lowry P. B., Meservy R. D., McDonald D. M. (2007). Genetic programming for prevention of cyberterrorism through dynamic and evolving intrusion detection. *Decision Support Systems*, vol. 43, n° 4, p. 1362–1374.
- Hassandoust F., Techatassanasoontorn A. A., Singh H. (2020). Information security behaviour: A critical review and research directions. In *European conference on information systems, ECIS 2020*.

- Johnston A. C., Warkentin M. (2010). Fear appeals and information security behaviors: an empirical study. *MIS quarterly*, p. 549–566.
- Kapp K. M. (2012). *The gamification of learning and instruction: game-based methods and strategies for training and education*. John Wiley & Sons.
- Kolb D. A. (1984). Experience as the source of learning and development. *Upper Sadle River: Prentice Hall*.
- Lallie H. S., Shepherd L. A., Nurse J. R. C., Erola A., Epiphaniou G., Maple C. *et al.* (2020, juin). Cyber security in the age of COVID-19: A timeline and analysis of cyber-crime and cyber-attacks during the pandemic. *ArXiv*, p. 1–20.
- Leach J. (2003). Improving user security behaviour. *Computers & Security*, vol. 22, n° 8, p. 685–692.
- Loch K. D., Carr H. H., Warkentin M. E. (1992). Threats to information systems: today's reality, yesterday's understanding. *Mis Quarterly*, vol. 16, n° 2, p. 173–186.
- Maaravi Y. (2020). Using hackathons to teach management consulting. *Innovations in Education and Teaching International*, vol. 57, n° 2, p. 220–230.
- McAlaney J., Benson V. (2020). Cybersecurity as a social phenomenon. In *Cyber influence and cognitive threats*, p. 1–8. Elsevier.
- Mirkovic J., Peterson P. A. H. (2014, août). Class capture-the-flag exercises. In *2014 USENIX summit on gaming, games, and gamification in security education (3gse 14)*. San Diego, CA, USENIX Association. Consulté sur <https://www.usenix.org/conference/3gse14/summit-program/presentation/mirkovic>
- Mitnick K., Simon W. (2003). *The art of deception: Controlling the human element of security*. John Wiley and Sons.
- Muscanel N. L., Guadagno R. E., Murphy S. (2014). Weapons of influence misused: A social influence analysis of why people fall prey to internet scams. *Social and Personality Psychology Compass*, vol. 8, n° 7, p. 388–396.
- Mustar P. (2009). Technology management education: Innovation and entrepreneurship at mines paristech, a leading french engineering school. *Academy of Management Learning & Education*, vol. 8, n° 3, p. 418–425.
- NIST. (2018, apr). *Framework for improving critical infrastructure cybersecurity, version 1.1*. Rapport technique. National Institute of Standards and Technology.
- Prekaj B., Stilo G., Madeddu L. (2020). Challenges and solutions to the student dropout prediction problem in online courses. In *Proceedings of the 29th acm international conference on information & knowledge management*, p. 3513–3514.
- Rath J., Ischi M., Perkins D. (2014). Evolution of different dual-use concepts in international and national law and its implications on research ethics and governance. *Science and engineering ethics*, vol. 20, n° 3, p. 769–790.
- Razmerita L., Kirchner K., Hockerts K., Tan C.-W. (2019, 12). Modeling collaborative intentions and behavior in digital environments: The case of a massive open online course (mooc). *Academy of Management Learning & Education*.

- Reix R. (2000). *Systemes d'information et management des organisations*. Paris, Vuibert.
- Sagarin B. J., Mitnick K. D. (2012). The path of least resistance. *Six Degrees Of Social Influence: Science, Application, and the Psychology of Robert Cialdini*, p. 12.
- Shah M. H., Jones P., Choudrie J. (2019). Cybercrimes prevention: promising organisational practices. *Information Technology & People*.
- Shropshire J. (2009). A canonical analysis of intentional information security breaches by insiders. *Information Management and Computer Security*, vol. 17, n° 4, p. 221-234.
- Siala H., Kutsch E., Jagger S. (2019). Cultural influences moderating learners' adoption of serious 3d games for managerial learning. *Information Technology & People*.
- Silic M., Lowry P. B. (2020, jan). Using design-science based gamification to improve organizational security training and compliance. *Journal of Management Information Systems*, vol. 37, n° 1, p. 129-161.
- Snyder J. (2018). *A framework and exploration of a cybersecurity education escape room*. Thèse de doctorat non publiée, Brigham Young University.
- Stanton J., Stam K., Mastrangelo P., Jolton J. (2005). Analysis of end user security behaviors. *Computers and Security*, vol. 24, n° 2, p. 124-133.
- Strom B. E., Applebaum A., Miller D. P., Nickels K. C., Pennington A. G., Thomas C. B. (2018, juillet). *MITRE ATT&CK: Design and philosophy*. Rapport technique. MITRE CORP BEDFORD MA.
- Thomson G. (2012, feb). BYOD: enabling the chaos. *Network Security*, vol. 2012, n° 2, p. 5-8.
- Tobias S., Fletcher J. D., Wind A. P. (2014). Game-based learning. In *Handbook of research on educational communications and technology*, p. 485-503. Springer.
- Tsohou A., Karyda M., Kokolakis S. (2015). Analyzing the role of cognitive and cultural biases in the internalization of information security policies: Recommendations for information security awareness programs. *Computers & security*, vol. 52, p. 128-141.
- Venkatesh V., Morris M. G., Davis G. B., Davis F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS quarterly*, p. 425-478.
- Ward R., Beyer B. (2014). Beyondcorp: A new approach to enterprise security. *login*, vol. 39, n° 6.
- Warkentin M., Willison R. (2009). Behavioral and policy issues in information systems security: the insider threat. *European Journal of Information Systems*, vol. 18, n° 2, p. 101-105.
- Willison R., Warkentin M. (2013). Beyond deterrence: an expanded view of employee computer abuse. *MIS Quartely*, vol. 37, n° 1, p. 1-20.
- Wu F., Narang H., Clarke D. (2014). An overview of mobile malware and solutions. *Journal of Computer and Communications*, vol. 02, n° 12, p. 8-17.
- Zhi-Jun W., Hai-Tao Z., Ming-Hua W., Bao-Song P. (2012). MSABMS-based approach of detecting Idos attack. *Computers & Security*, vol. 31, n° 4, p. 402-417.



---

## Détection de fraude financière dans un système de transactions interbancaires.

**Hamza Chergui**<sup>1,2</sup>, **Lylia Abrouk**<sup>1</sup>, **Nadine Cullot**<sup>1</sup>,  
**Nicolas Cabioch**<sup>2</sup>

1. Laboratoire d'Informatique de Bourgogne - EA 7534

Univ. Bourgogne Franche-Comté

9, avenue Alain Savary, F-21078 Dijon - France

hamza.chergui@etu.u-bourgogne.fr

lylia.gouaich-abrouk,nadine.cullot@u-bourgogne.fr

2. SKAIZen Group

14, rue de Mantes, 92700 Colombes

hchergui, ncabioch@skaizengroup.fr

---

*RÉSUMÉ. Les activités liées au blanchiment d'argent sont de plus en plus courantes. Les institutions financières se doivent d'améliorer leurs systèmes actuels jugés trop peu efficaces. Nous travaillons dans le domaine bancaire où nous analysons des transactions interbancaires internationales aux caractéristiques spécifiques d'un réseau nommé SWIFT. À travers une étude de l'état de l'art, nous montrons les limites des techniques existantes d'apprentissage automatique sur les caractéristiques de ces transactions. Nous proposons ainsi une approche de détection de transactions frauduleuses en deux étapes : (i) nous calculons de nouvelles caractéristiques spécifiques au réseau SWIFT, (ii) nous proposons une adaptation de l'algorithme de classification supervisé des K plus proches voisins (KNN) pour la détection d'anomalies. Une comparaison de nos travaux avec d'autres classifieurs sur un jeu de données réel montre l'apport et l'efficacité de notre approche.*

*ABSTRACT. The number of money laundering activities are increasing. Financial institutions must improve their systems judged not enough efficient. Our work is to analyse international and interbank transactions circulating in a network from a company named SWIFT. Through a study of the state-of-the-art, we expose the limits of the current machine learning techniques for the SWIFT transactions characteristics. We propose a two step fraud detection approach : (i) we compute new features related to SWIFT transactions, (ii) we propose a new metric of distance based on euclidean distance for KNN algorithm. A comparison of our work with other classifiers using a real data set prove the efficiency of our approach.*

*MOTS-CLÉS : apprentissage automatique, détection de fraude, finance*

*KEYWORDS: machine learning, fraud detection, finance*

---

## 1. Introduction

La lutte contre le blanchiment d'argent est une tâche complexe pour les institutions financières. (Jensen, 1997) définit le blanchiment d'argent comme une activité criminelle consistant à dissimuler l'origine ou le propriétaire de fonds obtenus de manière illégale dans le but de rendre ces fonds légitimes. Les institutions financières, principalement les banques, possèdent des systèmes de lutte contre le blanchiment d'argent, cependant ces systèmes ne sont pas assez efficaces. Selon (Knobel, 2019) 98,9% des activités de blanchiment d'argent passent à travers les mailles du filet. Les institutions financières se doivent d'améliorer leurs systèmes sous peine de recevoir des amendes conséquentes par les régulateurs du monde financier.

Les systèmes doivent être en mesure d'analyser des transactions financières qui sont des envois d'argent entre un émetteur et un bénéficiaire. Ces transactions peuvent être internationales et interbancaires. En outre, une transaction est caractérisée par des informations telles qu'un montant, une date ou une devise. Les systèmes actuels reposent sur des listes de règles et de surveillances, ce type de système est limité car les règles peuvent être identifiées par les criminels qui adaptent alors leurs manières de frauder rendant ainsi les systèmes inefficaces.

Le blanchiment d'argent est complexe à détecter du fait qu'il peut être effectué de multiples façons. (Schott, 2006) présente plusieurs schémas de blanchiment, comme la réalisation de multiples transactions de petits montants sur des comptes anodins, ou l'achat et la vente de biens luxueux tels que des voitures ou des bijoux. De plus, les recherches sur le sujet sont mises en difficulté par l'obtention de données qui sont privées. Notre travail s'inscrit dans les travaux de recherche de l'entreprise SKAIZen Group qui vise à améliorer la détection de fraude avec des données provenant d'un réseau d'une entreprise appelé SWIFT<sup>1</sup>. Il s'agit d'une entreprise qui met à disposition un réseau interbancaire qui propose différents services comme le transfert d'argent entre différents comptes bancaires. Ce réseau permet de réaliser des transactions financières entre plus de 11000 organismes bancaires à travers près de 200 pays. Parmi ces données transactionnelles peuvent se cacher des anomalies pouvant être liées à des activités de blanchiment d'argent. Ainsi, l'analyse de transactions interbancaires est un enjeu primordial pour les institutions financières. L'apprentissage automatique est un domaine de l'intelligence artificielle qui permet d'apprendre automatiquement à partir de données et d'améliorer les résultats grâce à une phase d'apprentissage. L'apprentissage automatique est utilisé dans plusieurs domaines comme la bourse pour prédire des séries temporelles ou analyser des profils d'utilisateurs dans le e-commerce. Actuellement, peu de travaux prennent en compte l'aspect international et les intermédiaires des transactions. Nous nous intéressons dans notre travail à l'utilisation et l'adaptation des techniques d'apprentissage automatique dans le domaine financier et plus particulièrement dans la détection de fraude dans des données transactionnelles internationales au format spécifique nommé SWIFT.

La suite de l'article est organisée de la manière suivante : dans la section 2 nous dres-

---

1. <https://www.swift.com/>

sons un état de l'art des techniques d'apprentissage automatique dans le domaine des fraudes financières et nous concluons cette section avec une synthèse. Après avoir expliqué les limites des techniques existantes dans la littérature pour nos types de données, nous proposons notre approche dans la section 3. Afin de la valider, nous avons réalisé des expérimentations à partir d'un jeu de données réel au sein de la section 4. Enfin, nous concluons et abordons nos perspectives de recherche dans la section 5.

## 2. État de l'art

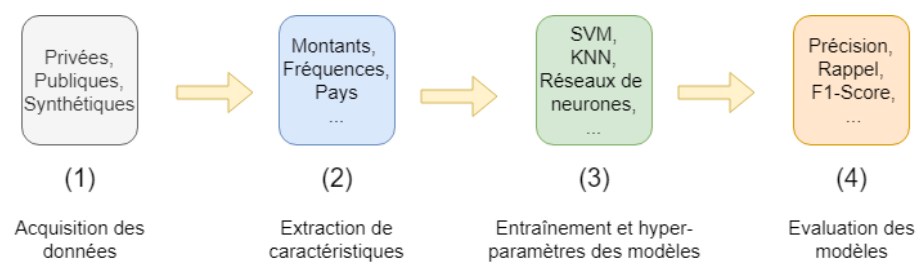


FIGURE 1. *Processus des techniques d'apprentissage automatique*

Selon (Mahesh, 2020), l'apprentissage automatique est utilisé pour apprendre aux machines à manipuler les données efficacement. L'objectif des techniques d'apprentissage automatique est d'apprendre à partir des données pour en extraire des informations. Dans cet article, nous avons dressé un état de l'art des techniques d'apprentissage automatique pour la détection de fraude financière qui suivent un processus pouvant se diviser en quatre étapes. Cette section est structurée selon ces quatre étapes : la première est **l'acquisition des données** qui peut être une tâche compliquée pour certains domaines tels que la finance ou la médecine où les données sont confidentielles. Certains travaux utilisent pour cela des données synthétiques afin de réaliser les expérimentations. **L'extraction de caractéristiques** consiste à calculer de nouvelles informations à partir des caractéristiques existantes et à réduire le nombre de dimensions tout en gardant l'information contenue dans l'ensemble des caractéristiques de départ. Le **choix des algorithmes et de leurs hyper-paramètres** dépend du nombre de dimensions, du volume et de la nature des données. Enfin, dans la dernière étape **l'évaluation**, les modèles prédictifs établis à l'aide des algorithmes sont évalués avec des mesures adaptées aux algorithmes.

### 2.1. L'acquisition des données

Les données financières sont connues pour être confidentielles, le manque de données publiques freine les expérimentations notamment pour leurs validations et leurs comparaisons. Il existe différents formats de données selon la source des données (banque de détail, banque agricole), les volumes de données des expérimentations de la littérature sont très hétérogènes allant du millier de données jusqu'aux millions.

Dans ce contexte, la communauté scientifique s'est intéressée à la génération de données synthétiques dans le domaine de la finance. (Lopez-Rojas, Axelsson, 2012) a développé un outil (PaySim) de simulateur de paiements mobiles comprenant des paiements frauduleux. (Michalak, Korczak, 2011) détaille la manière dont il a généré ses données pour valider ses expérimentations, ce qui n'est pas le cas de nombreuses expérimentations ayant recours à des données synthétiques (Tang, Yin, 2005 ; Keyan, Tingting, 2011).

Une technique exploitée est l'utilisation de données réelles couplée à des données frauduleuses générées artificiellement. Un jeu de données financier est un ensemble de transactions avec des attributs de base comme le montant et la date. Un jeu de données de référence utilisé dans de nombreuses expérimentations de notre domaine est un ensemble de transactions réalisées avec des cartes de crédits contenant des fraudes, le jeu de données comprend 3 attributs connus qui sont la date, le montant et la classe de la transaction (frauduleuse ou non frauduleuse) et 28 autres attributs dont la signification n'est pas connue<sup>2</sup>.

Les techniques d'apprentissage supervisé requièrent des données labellisées, les transactions ont un label qui précise si cette dernière est frauduleuse. La particularité de ces jeux de données est le déséquilibre de classe entre les transactions frauduleuses et les non frauduleuses, le ratio de transactions frauduleuses avoisine habituellement 1%.

## 2.2. Extraction de caractéristiques

Les données nécessitent d'être traitées avant d'alimenter les algorithmes. De nombreux types de traitement existent : la normalisation, l'ajout de nouvelles caractéristiques ou la réduction de dimension.

Les données financières présentes dans la littérature sont des ensembles de transactions avec des attributs de base comme le montant, la date et les acteurs impliqués (Kumar *et al.*, 2020 ; Porwal, Mukund, 2018 ; Carcillo *et al.*, 2021). Les caractéristiques calculées à partir de ces attributs sont des informations sur les clients extraites à partir de leur historique de transactions. Des agrégats sont réalisés par période (hebdomadaire, mensuel, annuel) afin de calculer des caractéristiques sur les montants moyens des transactions réalisées par les clients ainsi que leurs fréquences (nombre de transactions faites selon une période). La plupart des approches visent à détecter des comportements frauduleux avec des transactions ayant des sommes ou des fréquences trop élevées.

De plus, le problème de déséquilibre des classes peut être moins contraignant à l'aide de techniques comme SMOTE (Chawla *et al.*, 2002) permettant de générer des fraudes supplémentaires à partir des fraudes déjà existantes dans le jeu de données. (Badal-Valero *et al.*, 2018) prouvent l'efficacité de ce type de technique dans le domaine de la détection de fraude financière, en obtenant des meilleurs résultats en utilisant l'algorithme SMOTE.

2. <https://www.kaggle.com/mlg-ulb/creditcardfraud>

Les jeux de données ayant un nombre élevé de caractéristiques peuvent être problématiques car cela augmente les temps d'entraînement. Pour les techniques de réduction de dimensions, (Bestami Yuksel *et al.*, 2020) utilisent l'algorithme PCA pour entraîner l'algorithme des K plus proches voisins et (Paula *et al.*, 2016) utilisent les auto-encodeurs (technique d'apprentissage profond) pour réduire le nombre de dimensions du jeu de données, et obtiennent un temps d'entraînement 20 fois plus rapide.

### 2.3. *Choix des algorithmes et de leurs hyper-paramètres*

Au sein de la littérature, de nombreux travaux se concentrent sur la comparaison d'algorithmes d'apprentissage supervisé pour déduire quel algorithme est le mieux adapté aux données et à leurs volumes (Zhang, Trubey, 2019; Lorenz *et al.*, 2020; Bestami Yuksel *et al.*, 2020).

Des auteurs proposent des nouvelles architectures (Alarab *et al.*, 2020) ou des versions améliorées d'algorithmes comme le random forest avec (Xuan *et al.*, 2018).

Des algorithmes d'apprentissage non supervisé sont utilisés pour la détection de fraude financière, comme l'approche de (Porwal, Mukund, 2018) proposant une approche basée sur l'algorithme K-means, tandis que d'autres travaux se concentrent sur la proposition de nouvelles mesures de distance pour détecter des outliers (Larik, Haider, 2011).

Les techniques d'apprentissage semi supervisé consiste à utiliser dans un premier temps les techniques d'apprentissage non supervisé pour labeliser les données avec les clusters ou outliers, puis ces données labellisées alimentent un algorithme d'apprentissage supervisé (Le Khac, Kechadi, 2010; Raza, Haider, 2011; Carcillo *et al.*, 2021).

### 2.4. *Évaluation des modèles*

Une fois les modèles entraînés, l'objectif est de les évaluer afin de vérifier leur efficacité, pour cela de nombreuses mesures existent telles que la précision, le rappel ou le F1-Score. Ces mesures sont utilisées dans la grande majorité des travaux d'apprentissage automatique et permettent la comparaison des travaux entre eux. Cependant, il faut prendre en compte les volumes de données et les proportions des classes car celles-ci sont importantes et peuvent ne pas apparaître dans les mesures. Enfin, il est plus difficile d'évaluer des modèles avec des jeux de données non labellisés car il est compliqué de vérifier les résultats des prédictions. Il faut avoir recours à une vérification manuelle qui est une opération pouvant être longue et imprécise.

### 2.5. *Synthèse*

Les travaux présentés dans l'état de l'art des techniques d'apprentissage automatique dans le domaine de la détection de fraude financière montrent des verrous pour chacune des étapes. La difficulté d'acquisition de données financières freine la comparaison et la validation des approches, l'apport des données synthétiques peut

être remis en cause par leur potentiel écart avec la réalité. Concernant l'extraction de connaissances, les travaux se focalisent sur les caractéristiques portées sur les montants et fréquences des transactions des clients, on observe un manque de travaux sur les transactions comportant des aspects internationaux et interbancaires. Pour le choix des algorithmes, il existe de nombreux travaux comparant les différents algorithmes d'apprentissage automatique. Ces derniers nous ont permis de sélectionner les plus utilisés pour notre étude. Nous avons retenu l'algorithme des K plus proches voisins en l'adaptant pour les transactions SWIFT. Ce choix s'explique par l'utilisation dans l'algorithme d'une distance adaptée aux transactions SWIFT. Cette distance nous permet d'exploiter les attributs des pays et devises des transactions et donc de proposer une solution pour la détection de fraude financière sur des transactions possédant une dimension internationale. De plus, à notre connaissance, il n'existe pas de travaux opérant avec des données du réseau SWIFT hormis l'article de (Alkhalili *et al.*, 2021) qui utilise les données SWIFT pour vérifier les transactions qui ont été bloquées car les clients figuraient dans des listes de surveillances.

En résumé, les approches présentées dans la littérature ne répondent pas aux besoins de nos données financières possédant des dimensions internationales et interbancaires, c'est pourquoi nous proposons une approche pour la détection de fraude financière sur des données SWIFT.

*EC = Extraction de caractéristiques*

*RD = Réduction de dimensionnalité*

*SVM = Machine à vecteurs de support (Single Vector Machine)*

*NB = Classification Naïve bayésienne*

*RF = Forêt aléatoire (Random Forest)*

*KNN = K plus proches voisins (K-Nearest Neighbors)*

TABLEAU 1. *Tableau comparatif des travaux de la littérature selon des critères d'analyse*

Référence (année)	Données	Algorithme	EC	RD	Types de fraudes
(Tang, Yin, 2005) (2005)	Privée	SVM	✓		Montant, Fréquence
(Lv <i>et al.</i> , 2008) (2008)	Privée	Réseau de neurones	✓		Montant, Fréquence
(Le Khac, Kechadi, 2010) (2010)	Privée	Semi-supervisé	✓		Montant, Fréquence
(Kumar <i>et al.</i> , 2020) (2018)	PaySim	Réseau Bayésien		✓	Montant
(Bestami Yuksel <i>et al.</i> , 2020) (2020)	Carte de credit	KNN		✓	Fréquence
(Alkhalili <i>et al.</i> , 2021) (2021)	SWIFT	SVM,NB,RF	✓		Montant, Pays
Notre approche	Privée	KNN	✓	✓	Montant, Fréquence, Pays, Devise

### 3. Approche

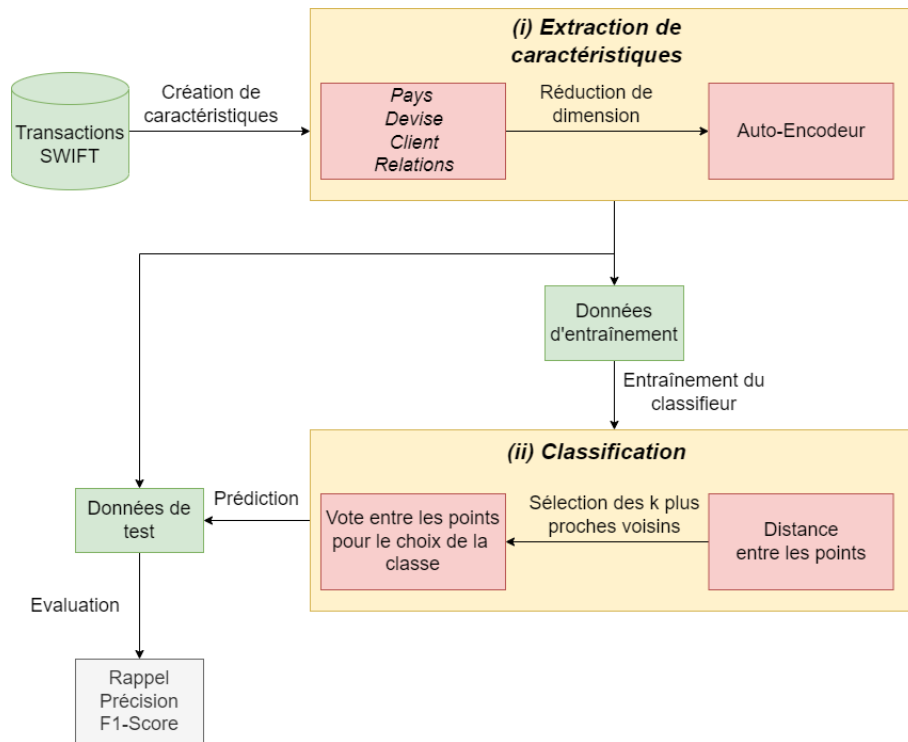


FIGURE 2. Schéma récapitulatif de l'approche

Nous proposons une approche permettant d'entraîner un classifieur capable d'apprendre sur un jeu de données transactionnelles labelisées. L'apprentissage prédira la classe (frauduleuse ou non frauduleuse) des transactions à partir de leurs caractéristiques.

Ce travail se déroule en deux étapes, dans un premier temps nous travaillons sur le jeu de données pour créer des nouvelles caractéristiques spécifiques aux transactions SWIFT notamment sur leur dimension internationale et interbancaire. Ensuite, nous verrons comment réduire le nombre élevé de caractéristiques avec une technique d'apprentissage profond : les auto-encodeurs. Lorsque notre jeu de données disposera d'un nombre satisfaisant de caractéristiques nous pourrons entraîner un classifieur. Nous nous sommes intéressés à l'algorithme des K plus proches voisins (KNN) que nous avons adapté aux transactions SWIFT en changeant la formule de distance.

Les transactions SWIFT ont un format spécial appelé ISO 150022 MT qui possède plusieurs types de messages MT associés à différents services. Pour nos travaux, nous nous sommes focalisés sur messages MT103 représentant un type de message utilisé par les institutions financières pour des virements d'argent internationaux.

Un message MT103 possède de nombreux champs<sup>3</sup> dont certains sont obligatoires et d'autres facultatifs. Nous avons basé nos travaux sur un jeu de données de MT103 avec un nombre de champs réduit.

### 3.1. Extraction de caractéristiques

TABLEAU 2. Exemple de messages SWIFT

Émetteur	Intermédiaire	Bénéficiaire	Date	Devise	Montant
BIC0FR01	BIC0IT01	BIC0FR02	210625	EUR	15006
BIC0US03	-	BIC0GB01	210625	GBP	33065
BIC0FR04	BIC0FR06	BIC0FR05	210626	EUR	100325

Sur le tableau 2 nous avons listé les attributs SWIFT sur lesquels nous allons baser nos travaux, en étudiant ces attributs nous pouvons observer 3 acteurs : l'émetteur, l'intermédiaire et le bénéficiaire. La transaction correspond à un transfert d'argent d'un montant d'argent sur une devise réalisé à une date entre l'émetteur et le bénéficiaire. Les transactions SWIFT ont des circuits adaptés aux relations entre la banque émettrice et bénéficiaire. Si ces dernières n'ont pas de relations directes alors la transaction passera par l'intermédiaire, sinon le circuit comporte seulement l'émetteur et le bénéficiaire. Les acteurs sont identifiés par un BIC, il s'agit d'un code dont nous pouvons extraire le pays de l'institution financière.

A partir de ces attributs, nous avons calculé des caractéristiques selon différents historiques et périodes :

*Caractéristiques* : Nous avons choisi de retenir les caractéristiques communes aux travaux présents dans l'état de l'art c'est-à-dire celles liées aux montants et aux fréquences : montant minimum, montant maximum, montant moyen, somme des montants, latence (durée depuis la dernière transaction), le nombre de transactions réalisées, le nombre de transactions réalisées par les pays (émetteurs, intermédiaires et bénéficiaires), le nombre de transactions réalisées avec la devise de la transaction, le nombre d'intermédiaires de la transaction. Puis à l'aide d'experts métiers, nous avons sélectionné de nouvelles caractéristiques spécifiques aux transactions SWIFT telles que le nombre de transactions réalisées vers les pays acteurs ou sur les devises et des caractéristiques sur la présence ou non d'un intermédiaire.

*Historique de transactions* : Nous avons calculé ces caractéristiques sur différents historiques de transactions, notamment avec les différentes combinaisons possibles entre les acteurs (émetteurs, intermédiaires et bénéficiaires) dans le but d'extraire des caractéristiques concernant leurs relations.

3. <http://www.iotafinance.com/SWIFT-ISO15022-Type-de-message-MT103.html>



*Période* : Enfin, nous avons étudié ces caractéristiques sur 3 périodes différentes (globale, mensuelle et décade). Les variations de comportements des acteurs ou de leurs relations peuvent être des signaux de fraudes. Cependant, ces variations peuvent également s’observer sur des ensembles d’acteurs partageant la même devise ou secteur d’activité, à cause d’évènements économiques pouvant avoir des impacts sur des périodes réduites (mensuelle ou décade).

### 3.2. Réduction de caractéristiques

Après ce travail d’extraction de caractéristiques, nous avons constaté un nombre élevé de dimensions sur notre jeu de données, ce qui va allonger considérablement le temps d’entraînement des classifieurs ce qui est contraignant pour la partie expérimentale. Ainsi, nous avons décidé d’utiliser des techniques de réduction de dimensions.

Nous avons fait le choix d’utiliser un auto-encodeur qui est un réseau de neurones ayant pour objectif d’apprendre une représentation d’un ensemble de données, on parle de représentation latente. Pour utiliser un auto-encodeur, il faut définir le nombre de couches et de neurones de l’encodeur et du décodeur. L’encodeur va permettre de réduire le nombre de dimensions des transactions, le décodeur lui va essayer de reformer la transaction de base en effectuant le travail inverse.

Dans l’exemple ci-dessous, on peut définir  $x$  comme une transaction, après avoir été transformée par l’encodeur la transaction est représentée par  $z$ , enfin après le décodeur la transaction sera représentée par  $\hat{x}$  qui doit avoir le même nombre de dimensions que  $x$ . Pour évaluer l’apprentissage de l’auto encodeur, on compare les représentations  $x$  et  $\hat{x}$ , notamment avec l’erreur quadratique moyenne. Les expérimentations seront réalisées sur la représentation latente des transactions ( $z$ ).

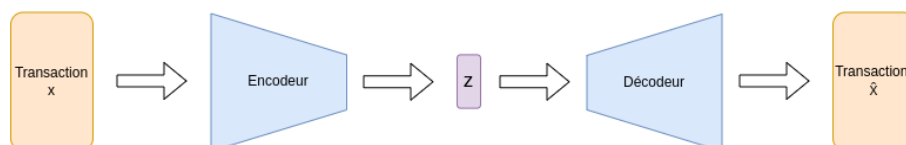
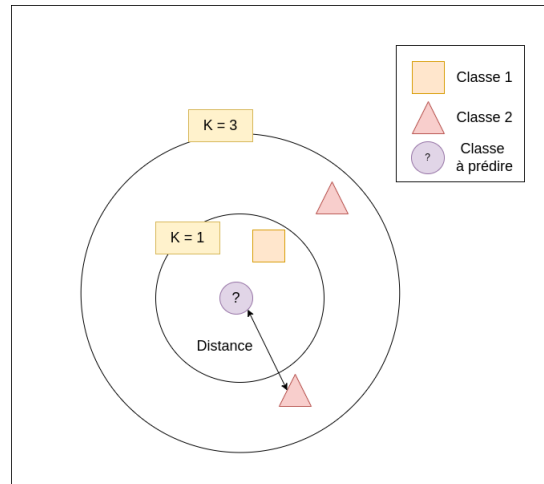


FIGURE 3. Schéma d’un auto-encodeur

### 3.3. Nouvelle mesure de distance pour l’algorithme KNN

Nous avons adapté l’algorithme KNN aux transactions SWIFT. Pour rappel, l’algorithme KNN fonctionne de la manière suivante : nous commençons par calculer la distance entre le point (la transaction) à classer avec les autres points du jeu de données, puis en fonction de ses  $k$  plus proches voisins on attribue la classe la plus représentée parmi les  $k$  plus proches voisins.

Dans notre contexte, les attributs des pays émetteurs, intermédiaires et bénéficiaires ainsi que la devise jouent un rôle dans la détection d’activités de blanchiment d’argent sur les transactions circulant à travers le réseau SWIFT. Nous avons constaté que les

FIGURE 4. *Fonctionnement KNN*

transactions possédant les mêmes pays et devises peuvent avoir des comportements similaires, nous souhaitons alors rapprocher les transactions possédant les mêmes devises et pays. En effet, les transactions peuvent avoir des comportements propres à chaque devise et pays, les événements politiques et économiques peuvent entraîner des conséquences directes sur certains pays et devises. Par exemple, si une devise commence à s'effondrer comme cela peut être le cas en Russie dans le contexte actuel<sup>4</sup>, il faut traiter différemment les transactions dont la devise est le rouble russe et les transactions comportant des acteurs russes des autres transactions avec des devises et des acteurs de nationalité différentes. Pour cela, en considérant que nos transactions sont des vecteurs, si deux transactions possèdent la même devise ou des pays similaires, nous réduisons leur distance, dans l'autre cas on l'augmente. D'un point de vue spatiale, on regroupe les transactions pour créer des groupes avec des pays et des devises similaires.

Le calcul de distance entre deux transactions est présenté sous forme d'un algorithme ci-dessous, l'avantage de notre calcul de distance est l'utilisation de variables catégoriques. En effet, il est difficile d'associer une valeur numérique à un pays ou une devise, les algorithmes d'apprentissage automatique fonctionnent majoritairement avec des valeurs numériques. C'est pourquoi inclure ces attributs pouvant avoir un impact sur des fraudes financières est important mais c'est une pratique absente de la littérature. Nous avons réalisé des expérimentations pour fixer la valeur des seuils en fonction du F1-Score. Nous avons étudié la valeur et l'impact de chaque seuil indépendamment.

4. <https://www.latribune.fr/economie/international/ukraine-les-pays-occidentaux-debranchent-la-russie-du-reseau-interbancaire-swift-l-arme-nucleaire-economique-904955.html>

---

**Algorithm 1** TS-Euclidienne

---

```

Require: Transaction  $t1, t2$ 
 $dst \leftarrow distance(t1, t2)$ 
if  $t1.devise == t2.devise$  then
     $dst \leftarrow dst * seuil\_devise$ 
else
     $dst \leftarrow dst * (1 + seuil\_devise)$ 
end if
if  $t1.pays\_emetteur == t2.pays\_emetteur$  then
     $dst \leftarrow dst * seuil\_emetteur$ 
else
     $dst \leftarrow dst * (1 + seuil\_emetteur)$ 
end if
if  $t1.pays\_intermdiaire == t2.pays\_intermdiaire$  then
     $dst \leftarrow dst * seuil\_intermdiaire$ 
else
     $dst \leftarrow dst * (1 + seuil\_intermdiaire)$ 
end if
if  $t1.pays\_beneficiaire == t2.pays\_intermdiaire$  then
     $dst \leftarrow dst * seuil\_intermdiaire$ 
else
     $dst \leftarrow dst * (1 + seuil\_intermdiaire)$ 
end if

```

---

#### 4. Expérimentation

Les expérimentations ont été réalisées avec un jeu de données de 200 000 transactions provenant du réseau SWIFT. Il s'agit de messages MT103 dont on a extrait les attributs présents sur le tableau 2.

Nous avons utilisé la plateforme Jupyter<sup>5</sup> pour développer notre approche grâce à laquelle nous avons comparé plusieurs classifieurs à l'aide de la librairie Scikit-Learn. Nous avons utilisé un jeu de données réelles labelisées obtenu grâce à une collaboration avec l'entreprise SKAIZen Group. Comme expliqué dans l'approche, nous avons commencé nos expérimentations en ajoutant de nouvelles caractéristiques basées sur les historiques des différents acteurs et également par rapport aux pays et devises des transactions. Ces caractéristiques ont été calculées sous différentes granularités, nous donnant ainsi un nombre élevé de dimensions. Notre approche basée sur l'algorithme KNN est sensible au nombre de dimensions du jeu de données, une haute dimensionnalité rallonge de manière considérable le temps d'entraînement. Pour résoudre cela, nous avons utilisé un auto-encodeur dont la configuration a été choisie à l'aide d'expériences présentées sur le tableau 3, l'objectif est de minimiser la me-

---

5. <https://jupyter.org/>

sure de l'erreur quadratique moyenne symbolisant la perte d'information après à la réduction de dimensions. Pour établir le nombre de couches nous avons testé plusieurs configurations avec 2,3 et 4 couches, la configuration avec 3 couches obtient l'erreur quadratique moyenne la plus faible. Pour le nombre de neurones par couche, nous avons d'abord fixé le nombre de la première couche avec le nombre de caractéristiques d'une transaction après le processus d'extraction de caractéristiques. Pour les deux autres couches, nous avons testé plusieurs configurations, nous avons retenu celle qui comporte 210 neurones sur la première couche, 150 sur la deuxième et 20 sur la troisième (210,150,20) qui possède l'erreur quadratique moyenne la plus faible (0.0065). A l'issue de l'étape de réduction de dimensions, nous avons obtenu un jeu de données comprenant 20 caractéristiques soit 10 fois moins que le nombre de dimensions obtenu après la phase d'extraction de caractéristiques.

TABLEAU 3. Résultats Auto-Encodeur

<i>Configuration</i>	<i>Erreur quadratique moyenne</i>
(210,150,80,20)	0.0077
(210,150,80,20)	0.0074
(210,100,20)	0.0071
(210,20)	0.0069
(210,150,20)	<b>0.0065</b>

Pour l'entraînement de notre modèle avec l'algorithme KNN, le jeu de données a été séparé de manière à avoir 80% du jeu de données en tant que jeu d'entraînement et 20% du jeu de données en tant que jeu de test. Nous avons fait le choix d'utiliser la distance euclidienne pour la disposition spatiale de nos données, nous avons expérimenté avec d'autres distances comme la distance Manhattan ou Minkowski, mais les expériences avec la distance euclidienne obtenaient les meilleurs résultats. Pour le choix des seuils de notre mesure de distance TS-Euclidean, nous avons réalisé différentes expérimentations en faisant varier leur valeur et en évaluant selon le F1-Score, ces expérimentations sont présentées sur la figure 5 qui montre que les valeurs 0.3, 0.6, 0.5 et 0.4 sont les meilleures pour les seuils respectifs de la devise, du pays émetteur, du pays intermédiaire et du pays bénéficiaire.

Afin d'évaluer nos modèles nous nous sommes basés sur les mesures suivantes :

- La précision : ratio entre le nombre d'instances correctement prédites d'une classe par rapport au nombre d'instances de la classe.
- Le rappel : ratio entre le nombre d'instances correctement prédites d'une classe par rapport au nombre d'instances prédites de la classe.
- Le F1-Score : moyenne pondérée entre la précision et le rappel.

Nous avons décidé de comparer notre classifieur KNN-TS Euclidean avec réduction de caractéristiques avec un auto-encodeur avec les classifieurs présents dans la littérature (SVM, Naive Bayes et Random Forest). Nous avons également notre mesure de distance TS Euclidean avec d'autres distances (Euclidienne, Manhattan et Min-

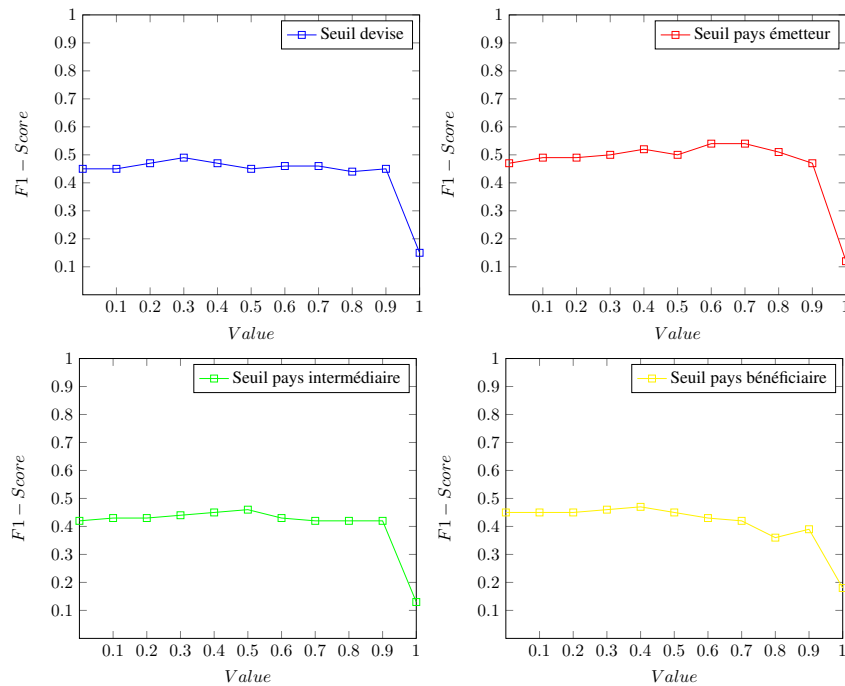


FIGURE 5. Résultats des expérimentations sur les seuils

kowski). Enfin, nous avons comparé l'utilisation de l'auto-encodeur avec l'algorithme d'analyse en composantes principales (ACP). Tous ces résultats sont présentés sur le tableau 3.

TABLEAU 4. Résultats des prédictions sur le jeu de test.

	Transactions Frauduleuses		
	Précision	Rappel	F1-Score
SVM	0.73	0.19	0.13
Naive Bayes	0.02	0.40	0.03
MLP	0.55	0.23	0.32
Random Forest	<b>0.85</b>	0.36	0.50
KNN-TS Euclidean (ACP)	0.61	0.23	0.33
KNN Manhattan (Auto-encodeur)	0.59	0.43	0.50
KNN Minkowski (Auto-encodeur)	0.59	0.43	0.50
KNN Euclidean (Auto-encodeur)	0.62	0.44	0.51
KNN-TS Euclidean (Auto-encodeur)	0.65	<b>0.48</b>	<b>0.55</b>

Les résultats sont présentés dans le tableau 4. Lors de nos expériences nous avons observé que les résultats étaient conformes à nos attentes. Nous obtenons un meilleur rappel (0.48) et F1-Score (0.55) avec notre approche, cependant l'algorithme Random

Forest dépasse notre modèle en terme de précision. La précision d'un modèle prédictif dans la détection de fraude permet d'alléger le travail des experts qui doivent vérifier si les transactions détectées comme frauduleuses sont de fausses alertes. Quant à l'amélioration du rappel, il s'agit d'améliorer la détection de transactions frauduleuses, un bon rappel signifie qu'il y a peu de transactions frauduleuses non détectées. A travers nos expérimentations, pour la réduction de dimension nous montrons l'apport des auto-encodeurs avec une précision, un rappel et un F1-Score supérieur à un modèle utilisant les ACP. Nous comparons également trois types de distances (Euclidienne, Manhattan et Minkowski), les résultats montrent que la distance euclidienne donne les meilleurs résultats.

## 5. Conclusion

En conclusion, nous avons dressé un état de l'art des techniques d'apprentissage automatique pour la détection de fraude financière. Nous avons détaillé le processus de ces techniques se décomposant en quatre étapes, nous avons réalisé un parallèle avec les caractéristiques de nos données : les transactions SWIFT. Leurs dimensions internationales et interbancaires sont importantes et peuvent avoir un impact sur la détection de fraude financière. Nous montrons cela à travers notre approche basée sur les techniques d'apprentissage supervisé. Nous commençons par calculer des nouvelles caractéristiques à partir de celles présentes dans la littérature et nous en sélectionnons de nouvelles spécifiques aux transactions SWIFT. Ensuite, nous avons utilisé les auto-encodeurs pour réduire le nombre élevé de caractéristiques obtenues pour enfin entraîner le classifieur KNN. Pour cet algorithme nous avons introduit une nouvelle mesure de distance basée sur les devises et les pays présents dans les transactions. Les résultats nous permettent de valider notre approche et notre hypothèse selon laquelle les dimensions internationales des transactions SWIFT ont un impact sur la détection des fraudes financières. Pour nos travaux futurs, nous souhaitons étendre nos approches à d'autres techniques notamment basées graphes et des ontologies. Nous aimerions également tester notre approche au sein d'institutions financières.

**Remerciement** Ce travail est soutenu à la fois par l'entreprise SKAIZen Group et l'ANRT.

## Bibliographie

- Alarab I., Prakoonwit S., Nacer M. I. (2020). Competence of graph convolutional networks for anti-money laundering in bitcoin blockchain. In *Proceedings of the 2020 5th international conference on machine learning technologies*, p. 23–27.
- Alkhalili M., Qutqut M. H., Almasalha F. (2021). Investigation of applying machine learning for watch-list filtering in anti-money laundering. *IEEE Access*, vol. 9, p. 18481–18496.
- Badal-Valero E., Alvarez-Jareño J. A., Pavía J. M. (2018). Combining benford's law and machine learning to detect money laundering. an actual spanish court case. *Forensic science international*, vol. 282, p. 24–34.

- Bestami Yuksel B., Bahtiyar S., Yilmazer A. (2020). Credit card fraud detection with nca dimensionality reduction. In *13th international conference on security of information and networks*, p. 1–7.
- Carcillo F., Le Borgne Y.-A., Caelen O., Kessaci Y., Oblé F., Bontempi G. (2021). Combining unsupervised and supervised learning in credit card fraud detection. *Information sciences*, vol. 557, p. 317–331.
- Chawla N. V., Bowyer K. W., Hall L. O., Kegelmeyer W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, vol. 16, p. 321–357.
- Jensen D. (1997). Prospective assessment of ai technologies for fraud detection: A case study. In *Aaai workshop on ai approaches to fraud detection and risk management*, p. 34–38.
- Keyan L., Tingting Y. (2011). An improved support-vector network model for anti-money laundering. In *2011 fifth international conference on management of e-commerce and e-government*, p. 193–196.
- Knobel A. (2019). *Swift data can be a global vantage point for tackling global money laundering*. Consulté sur <https://taxjustice.net/2019/07/11/swift-data-can-be-a-global-vantage-point-for-tackling-global-money-laundering/>
- Kumar A., Das S., Tyagi V. (2020). Anti money laundering detection using naïve bayes classifier. In *2020 ieee international conference on computing, power and communication technologies (gucon)*, p. 568–572.
- Larik A. S., Haider S. (2011). Clustering based anomalous transaction reporting. *Procedia Computer Science*, vol. 3, p. 606–610.
- Le Khac N. A., Kechadi M.-T. (2010). Application of data mining for anti-money laundering detection: A case study. In *2010 ieee international conference on data mining workshops*, p. 577–584.
- Lopez-Rojas E. A., Axelsson S. (2012). Money laundering detection using synthetic data. In *Annual workshop of the swedish artificial intelligence society (sais)*.
- Lorenz J., Silva M. I., Aparício D., Ascensão J. T., Bizarro P. (2020). Machine learning methods to detect money laundering in the bitcoin blockchain in the presence of label scarcity. In *Proceedings of the first acm international conference on ai in finance*, p. 1–8.
- Lv L.-T., Ji N., Zhang J.-L. (2008). A rbf neural network model for anti-money laundering. In *2008 international conference on wavelet analysis and pattern recognition*, vol. 1, p. 209–215.
- Mahesh B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR)*. [Internet], vol. 9, p. 381–386.
- Michalak K., Korczak J. (2011). Graph mining approach to suspicious transaction detection. In *2011 federated conference on computer science and information systems (fedcsis)*, p. 69–75.
- Paula E. L., Ladeira M., Carvalho R. N., Marzagao T. (2016). Deep learning anomaly detection as support fraud investigation in brazilian exports and anti-money laundering. In *2016 15th ieee international conference on machine learning and applications (icmla)*, p. 954–960.
- Porwal U., Mukund S. (2018). Credit card fraud detection in e-commerce: An outlier detection approach. *arXiv preprint arXiv:1811.02196*.

- Raza S., Haider S. (2011). Suspicious activity reporting using dynamic bayesian networks. *Procedia Computer Science*, vol. 3, p. 987–991.
- Schott P. A. (2006). *Reference guide to anti-money laundering and combating the financing of terrorism*. World Bank Publications.
- Tang J., Yin J. (2005). Developing an intelligent data discriminating system of anti-money laundering based on svm. In *2005 international conference on machine learning and cybernetics*, vol. 6, p. 3453–3457.
- Xuan S., Liu G., Li Z., Zheng L., Wang S., Jiang C. (2018). Random forest for credit card fraud detection. In *2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC)*, p. 1–6.
- Zhang Y., Trubey P. (2019). Machine learning and sampling scheme: An empirical study of money laundering detection. *Computational Economics*, vol. 54, n° 3, p. 1043–1063.



---

# Construction d'une ontologie dans le domaine financier pour la détection de fraudes

**Benjamin Auger<sup>1</sup>, Hamza Chergui<sup>1,2</sup>, Yara Chehade<sup>1</sup>,  
Jana El Kadri<sup>1</sup>, Lylia Abrouk<sup>1</sup>, Nicolas Cabioch<sup>2</sup>**

1. *Laboratoire d'Informatique de Bourgogne - EA 7534*

*Univ. Bourgogne Franche-Comté*

*9, Avenue Alain Savary, F-21078 Dijon - France*

*hamza.chergui,yara\_chehade,jana\_el-kadry@etu.u-bourgogne.fr*

*lylia.abrouk,benjamin.auger@u-bourgogne.fr*

2. *SKAIZen Group*

*14, rue de Mantes, 92700 Colombes*

*hchergui,ncabioch@skaizengroup.fr*

---

*RÉSUMÉ. L'objectif de notre travail est de peupler automatiquement à partir de sources hétérogènes une base de connaissances du domaine financier et structurer les informations extraites afin de permettre aux clients d'optimiser leurs moteurs de détection de fraude et de contrôles de sanction. Nous présentons dans cet article notre projet en cours de construction d'une ontologie financière basée sur la base de connaissances KYC et des transactions internationales interbancaires SWIFT aux caractéristiques spécifiques. Notre travail s'inscrit dans le cadre d'un projet de collaboration avec une entreprise dans le domaine financier.*

*ABSTRACT. The goal of our work is to automatically enrich a financial knowledge base from heterogeneous sources and structure the information extracted in order to allow clients to optimize their fraud detection and sanction control engines. We present in this article our project of financial ontology based on the KYC knowledge base and SWIFT international interbank transactions with specific characteristics. Our work is part of a collaborative project with a company in the financial domain.*

*MOTS-CLÉS : Ontologie, détection de fraude, finance, KYC*

*KEYWORDS: Ontology, fraud detection, finance, KYC*

---

## 1. Introduction

Dans le cadre des contrôles anti-fraude, la lutte contre le blanchiment d'argent ainsi que le financement du terrorisme, chaque institution financière a pour obligation

d'avoir des systèmes en place pour lutter contre ces activités illégales. Ces systèmes doivent pouvoir analyser des transactions financières pouvant avoir une dimension internationale et interbancaire. C'est dans ce contexte que l'entreprise SKAIZen Group développe un projet de recherche et d'innovation qui a pour but de construire une vision 360° des clients et des contreparties financières visant d'une part à peupler, à partir de sources hétérogènes, une base de faits vérifiés et d'autre part à permettre à leurs clients d'optimiser leurs moteurs de détection de fraudes et de contrôle de sanction associés aux transactions financières. La création d'une ontologie spécialisée sur les informations financières et liées à la conformité bancaire est une première étape dans ce projet de construction d'une base de connaissances. En organisant les données sous forme d'une ontologie spécifiée, nous pourrions interroger la base de connaissances sur les relations entre différentes entités (personnes ou organisations). Notre démarche de construction de base de connaissances consistera en un processus de peuplement d'une ontologie spécialisée dans les contrôles de flux financiers (conformité bancaire, KYC *Know Your Customer*, etc.).

Des ontologies spécialisées existent déjà dans divers domaines scientifiques et sont utilisées pour la description de l'information dans différentes spécialités (formation, biologie, médecine...). (Noy *et al.*, 2008) décrivent un portail d'ontologies biomédicales où (Psyché *et al.*, 2003) discute l'apport des ontologies dans les environnements de formation à distance. Dans la littérature, nous avons également identifié plusieurs ontologies orientées pour l'économie ou la finance. L'ontologie REA (Resource, Evènement, Agent) a été proposée initialement par McCarthy (McCarthy, 1982) pour les applications de comptabilité. Dans le modèle REA, les Agents comprennent les entités comme organisations ou individus impliquées dans un évènement qui décrit un type de relation économique. Cette ontologie est étendue *REA Business Management Ontology* (Schwaiger, 2016) pour l'entreprise et la gestion d'information.

Dans le domaine financier, la détection de fraudes est basée sur divers informations (transaction, profil client, ...). Ces dernières sont nombreuses, d'ordre qualitatif ou quantitatif et sont de sources hétérogènes, chez plusieurs acteurs du domaine (institutions financières, entreprises, ...). A notre connaissance, il n'existe pas d'ontologie décrivant des transactions bancaires dans le domaine financier et plus particulièrement des transactions interbancaires internationales aux caractéristiques particulières SWIFT. Par conséquent, la construction d'une telle ontologie nous permettrait de décrire ce modèle SWIFT qui vise à devenir une norme ISO 20022<sup>1</sup> et nous permettre de détecter des fraudes dans ces transactions grâce à la définition de règles métiers explicitant le statut des transactions en fonction des informations dans l'ontologie. Nous présentons dans la section suivante notre travail en cours de construction d'ontologie financière.

---

1. <https://www.iso20022.org/iso-20022-message-definitions>

## 2. L'ontologie financière CTO

L'objectif de notre projet est de construire une ontologie transactionnelle adaptée aux transactions SWIFT. La réalisation d'un tel projet aura un double apport pour notre entreprise car elle pourra se coupler notamment à la fois avec nos travaux sur l'extraction de connaissances d'articles financiers (Jabbari *et al.*, 2019 ; 2020) et de nos travaux sur la détection de fraude financière (Chergui *et al.*, 2022). En effet, à notre connaissance il n'existe pas de travaux couplant des approches d'apprentissage automatique et ontologique sur la détection de fraude financière. Nous avons pour ambition de proposer à la fois une démarche sur la construction d'une ontologie dans le domaine financier (SWIFT) et une approche sur la démarche d'implémenter une solution avec des techniques d'apprentissage automatique.

Les transactions SWIFT sont des messages répondant à une norme appelée ISO, cette dernière décrit les standards de formalisation de ces messages. Actuellement, ces messages suivent la norme ISO 150022 MT<sup>2</sup>, qui sera modifiée dans les prochaines années. En effet, cette norme migrera pour une nouvelle appelée ISO 20022 MX à partir de novembre 2022 où les messages auront la spécificité d'être formalisés à l'aide du langage XML. En outre, les messages MT correspondant à l'ancien format pourront encore être utilisés jusqu'en 2025.

Plusieurs types de messages MT ou MX sont utilisés pour des fonctions précises au sein du réseau SWIFT par lequel circule tous les messages. Pour nos travaux, nous avons décidé de nous intéresser aux virements d'argent internationaux représentés par le type "103" pour les messages MT (norme ISO 150022) et "pacs.008" pour les messages MX (norme ISO 20022). De nombreuses informations sur une transaction sont présentes dans ces messages. Nous avons sélectionné dans un premier temps les informations pertinentes dans le cadre de détection de fraude.

Nous avons choisi de retenir trois classes : la classe **Personne** correspond aux clients présents dans la base de connaissances KYC. La classe **Compte** correspond au compte d'un client bancaire des **personnes**, cette classe permet de faire la jonction entre les personnes présentes dans la base de connaissances et les personnes réalisant des transactions. La classe **Transaction** comporte les attributs montant, devise et date, une transaction est réalisée entre deux **comptes**. Le Tableau 1 montre les relations entre les concepts.

Nous définissons les contraintes de cardinalités suivantes :

- Une personne peut avoir plusieurs comptes bancaires  
 $Personne \subseteq \exists \text{ possède. Comptes}$
- Un compte n'appartient qu'à 1 seule personne  
 $Comptes \subseteq \doteq 1 \text{ inv\_possède. Possède}$
- Un compte effectue 1 à plusieurs transactions (transfert d'argent)  
 $Comptes \subseteq \exists \text{ effectue Transaction. Transactions}$

2. <https://help.flywire.com/hc/fr/articles/360012919014-Qu-est-ce-qu-un-message-SWIFT-MT103->

TABLEAU 1. Rôles avec leurs domaines

Rôle	Domain	Range	Rôle inverse
<b>possède</b>	Personne	Compte	inv_possède
<b>effectueTransaction</b>	Compte	Transaction	inv_effectueTransaction
<b>reçoitTransaction</b>	Compte	Transaction	inv_reçoitTransaction

– Un compte reçoit 1 à plusieurs transactions (reçu d’un transfert d’argent)  
 $\text{Compte} \subseteq \exists \text{reçoitTransaction.Transaction}$

Nous avons conçu une première version de notre ontologie sous l’outil protégé<sup>3</sup> avec les concepts présents sur la figure 1.

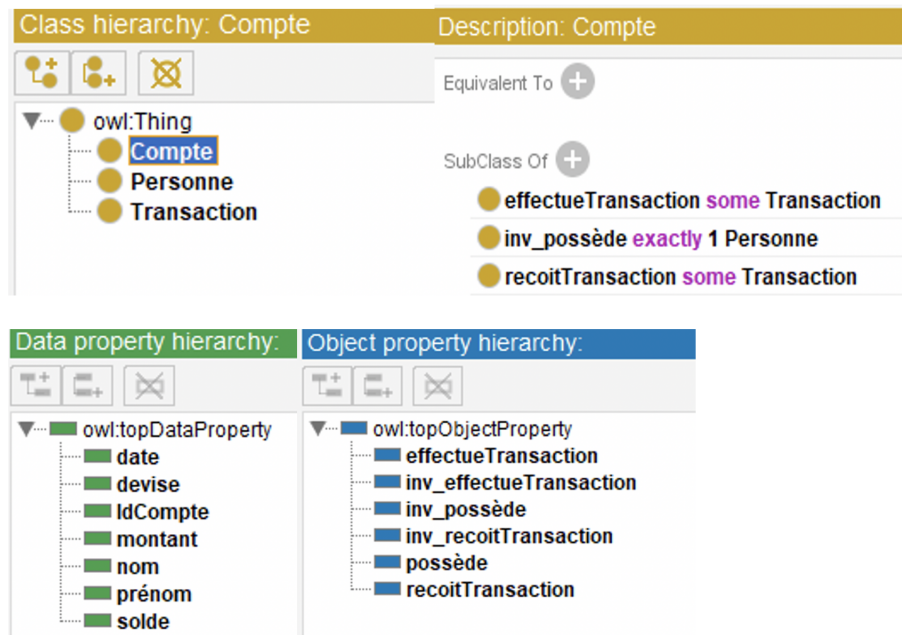


FIGURE 1. L’ontologie avec l’outil protégé

Une fois l’ontologie peuplée avec des messages MT et MX, nous avons rédigé quelques règles SWRL (Semantic Web Rule Language) pour la détection de fraudes. Par exemple la règle affichée ci-dessous permet d’obtenir les transactions réalisées

3. <https://protege.stanford.edu/>

de plus de 13000 dollars par un compte Iranien possédant un solde inférieur à 10000 dollars. Enfin, nous pourrions faire des recherches supplémentaires sur ces transactions à l'aide de notre base de connaissance.

```
Rules:
Rules +
effectueTransaction(?c, ?t), montant(?t, ?m), greaterThan(?m, 13000), devise(?t, ?d),
stringEqualIgnoreCase(?d, "USD"), solde(?c, ?s), lessThanOrEqual(?s, 10000), pays(?c, ?p),
stringEqualIgnoreCase(?p, "Iran") -> TransactionFrauduleuse(?t)
```

FIGURE 2. Exemple de règle de détection de fraude

L'efficacité de la proposition sera évaluée à l'aide d'un jeu de données réelles labellisées obtenu grâce à une collaboration avec l'entreprise SKAIZen Group. Notre ontologie sera enrichie à l'aide de ce jeu de données qui est composé d'un ensemble de transactions. Ces dernières possèdent un label qui nous permet de savoir s'il s'agit d'une transaction frauduleuse ou légitime. Une fois l'ontologie enrichie, nous pourrions faire tourner nos règles de détection de fraude qui sortiront les transactions jugées frauduleuses. Nous pourrions ensuite vérifier à l'aide des labels si les transactions frauduleuses le sont réellement. D'une manière plus globale, nous évaluerons notre proposition à partir des mesures de précision, de rappel et de F1-Score qui nous permettront de mesurer l'efficacité de la proposition.

### 3. Conclusion et perspectives

Dans le cadre de notre projet en cours, nous mettons en place un cadre formel et un outil permettant la détection de crimes financiers en modélisant le comportement des clients en utilisant les données transactionnelles. La construction de l'ontologie financière CTO est la première étape de ce travail. Dans la suite de ce projet, nous allons développer des règles métier complexes et adaptables en fonction de l'évolution des règles de fraudes et avec l'aide d'experts du domaine. Ce travail sera couplé par la suite avec notre base de connaissances KYC en enrichissant l'ontologie CTO avec le profil des clients à partir de sources de données hétérogènes.

**Remerciements** Ce travail est soutenu à la fois par l'entreprise SKAIZen Group, l'ANRT et l'ANR (France Relance).

### Bibliographie

- Chergui H., Alfred R. A., Abrouk L., Jabbari A., Cullot N. (2022). Détection d'anomalies: une méthode appliquée aux transactions interbancaires. *Extraction et Gestion des Connaissances: EGC'2022*, vol. 38.
- Jabbari A., Sauvage O., Cabioch N. (2019). Towards a knowledge base of financial relations: Overview and project description. In *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (Aike)*, p. 313–316.

- Jabbari A., Sauvage O., Zeine H., Chergui H. (2020). A french corpus and annotation schema for named entity recognition and relation extraction of financial news. In *Proceedings of the 12th language resources and evaluation conference*, p. 2293–2299.
- Mccarthy W. (1982, 07). The rea accounting model: A generalized framework for accounting systems shared data environment. *The Accounting Review*.
- Noy N. F., Shah N., Dai B., Dorf M., Griffith N., Jonquet C. *et al.* (2008). Bioportal: A web repository for biomedical ontologies and data resources. In C. Bizer, A. Joshi (Eds.), *Proceedings of the poster and demonstration session at the 7th international semantic web conference (iswc2008), karlsruhe, germany, october 28, 2008*, vol. 401. CEUR-WS.org. Consulté sur [http://ceur-ws.org/Vol-401/iswc2008pd\\_submission\\_25.pdf](http://ceur-ws.org/Vol-401/iswc2008pd_submission_25.pdf)
- Psyché V., Mendes O., Bourdeau J. (2003, 01). Apport de l'ingénierie ontologique aux environnements de formation à distance. *Sciences et Technologies de l'Information et de la Communication pour l'Éducation et la Formation*, vol. 10.
- Schwaiger W. S. A. (2016). Rea business management ontology: Conceptual modeling of accounting, finance and management control. In *Caise forum*.

---

# Evaluation de la valeur des données

## Modèle et méthode

**Jacky Akoka<sup>1</sup>, Isabelle Comyn-Wattiau<sup>2</sup>**

1. CEDRIC-CNAM & IMT Business School

2 rue Conté, Paris, FRANCE

[jacky.akoka@lecnam.net](mailto:jacky.akoka@lecnam.net)

2. ESSEC Business School

---

*RÉSUMÉ. Les données sont un actif stratégique pour les organisations. Évaluer leur valeur permet d'identifier les stratégies offrant un avantage concurrentiel et de mesurer le capital informationnel d'une entreprise. Il n'existe pas d'approche ni de normes et règles à appliquer dans cette évaluation du fait du caractère immatériel des données. Dans cet article, nous proposons une approche combinant un concept enrichi de valeur de la donnée, un modèle conceptuel incluant entre autres ce concept et la notion de contexte et une méthode d'évaluation de la valeur s'appuyant sur ce modèle. L'originalité est de combiner les approches fondées sur les risques avec les approches plus classiques d'évaluation des actifs tangibles (approche par les coûts, par le marché ou par le revenu). L'approche est confrontée à plusieurs cas réels issus de la littérature permettant une validation de son utilité et de sa valeur ajoutée.*

*ABSTRACT. Data is a strategic asset for organizations. Assessing its value provides companies with a mean to identify strategies that offer a competitive advantage and to measure their informational capital. There is neither an approach nor a set of standards and rules to apply for this valuation due to the intangible nature of data. In this paper, we propose an approach combining an enriched concept of data value, a conceptual model including this concept and the notion of context, and a value assessment method based on this model. The originality is to combine risk-based approaches with the more classical approaches to the valuation of tangible assets (cost, market, or income approaches). The approach is confronted with several real cases from the literature allowing a validation of its usefulness and added-value.*

*MOTS-CLES : Valeur des données, évaluation, modèle conceptuel, approche hiérarchique multicritères, contexte, risque, coût.*

*KEYWORDS: Data value, data valuation, conceptual model, analytical hierarchy process (AHP), context, risk, cost.*

---

## 1. Introduction

Les données jouent un rôle central dans le fonctionnement des organisations. Elles sont les actifs intangibles qui sous-tendent l'économie numérique. Les entreprises s'appuient sur les données pour prendre des décisions, notamment pour leurs investissements, ou pour générer des indicateurs de performance. Selon McKinsey, plus d'un tiers des ventes d'Amazon proviennent de son moteur de recommandation. Netflix affirme que ses recommandations fondées sur des algorithmes de traitement des données lui permettent d'économiser un milliard de dollars par an. La capitalisation boursière de Facebook est d'environ 1 000 milliards de dollars, mais la valeur nette de l'entreprise fondée sur l'actif et le passif n'est que de 138 milliards de dollars. La différence en termes de valeur s'explique par les données que Facebook collecte auprès des utilisateurs et qu'elle utilise à son tour pour alimenter ses algorithmes publicitaires. Les données sont donc un actif essentiel de nombreuses entreprises.

L'évaluation de la valeur des données est le processus qui consiste à attribuer une valeur monétaire qui représente l'importance économique des données pour l'organisation et ses différentes parties prenantes (employés, actionnaires, clients, fournisseurs, etc.). Toutefois, ce concept de valeur est multi facettes. Il s'agit bien sûr de valeur monétaire, mais aussi d'un concept économique relatif au bien-être social et sociétal. Dans ce cas, les données génèrent de la valeur lorsqu'elles permettent aux entreprises de créer des emplois, de devenir plus productives ou encore aux gouvernements de fournir des services publics plus efficaces.

L'évaluation de la valeur n'est pas simple. D'une part, les mêmes données peuvent être très importantes pour une organisation et n'avoir aucune valeur pour une autre. D'autre part, la maturité de certaines organisations ne leur permet pas de comprendre la valeur de leurs données et, encore moins, de l'évaluer. Enfin, le choix d'une méthode d'évaluation n'est pas aisé.

L'objectif de cet article est de proposer un modèle et une méthode permettant d'assister les entreprises dans le processus d'évaluation de la valeur de leurs données. Une contribution originale est la prise en compte combinée des dimensions *valeur, risque et coût* pour une meilleure appréhension de la valeur résultante. Nous proposons une approche d'évaluation de la valeur fondée sur trois piliers. Le premier est la définition du concept de *Valeur de la donnée*. Le second est un modèle conceptuel décrivant les informations nécessaires à l'évaluation de cette valeur des données. Le troisième pilier est une méthode d'aide à l'évaluation de la valeur de la donnée fondée sur le modèle conceptuel et structurée suivant l'approche hiérarchique multicritère.

Le reste de cet article est organisé comme suit. La deuxième section propose un état de l'art des principaux concepts et approches d'évaluation de la donnée. La



troisième section décrit successivement les trois piliers de l'approche et sa validation. Enfin, la quatrième section conclut et esquisse des voies de recherche future.

## **2. L'évaluation de la valeur des données : un état de l'art**

De nombreux articles décrivent les données comme des actifs intangibles (Savona, 2019; Corrado, 2019; Otto, 2015; Wdowin et Diepeveen, 2020). Ces actifs sont soumis à des réglementations et évalués à l'aide de techniques spécifiques (Ciuriak, 2019). De nature intangible, les données n'obéissent pas à la même dynamique concurrentielle que les actifs tangibles traditionnels (équipements, stocks, immobilier, etc.). Toutefois, leur gestion suit ou devrait suivre une approche similaire. En d'autres termes, il est impératif de comprendre la valeur des données et de les gérer tout au long de leur cycle de vie. Il est tout aussi important d'évaluer les risques liés à leur exploitation et de mesurer le retour sur investissement des projets d'acquisition et de transformation de données (Short et Todd, 2017).

Les données ont un coût parfois considéré comme irrécupérable (Wang et Zhao, 2020) ou dont le retour sur investissement est difficile à évaluer. En effet, la manière dont elles créent de la valeur n'est pas toujours claire pour les entreprises (Zeiter *et al.*, 2021). Elles peuvent notamment créer de la valeur en éclairant les décisions. Par exemple, elles peuvent permettre de concentrer les ressources limitées, dont disposent les organisations, sur les domaines qui créeront le plus de valeur (Attard et Brennan, 2018). Certains auteurs considèrent que la valeur des données augmente avec leur utilisation notamment en exploitant les métadonnées. D'autre part, plus les données sont précises, plus elles ont de la valeur (PwC, 2019). Toutefois, la qualité et la quantité des données sont soumises à la loi des rendements décroissants (Moody et Walsh, 1999). Certaines données sont périssables, par exemple, les données sur les clients qui ne sont pas régulièrement mises à jour. Enfin, plusieurs facteurs, comme l'exactitude, l'intégrité, la disponibilité et la fraîcheur, affectent le coût de la donnée.

Les gouvernements et les organismes internationaux mettent en place des initiatives pour mieux maîtriser l'actif « donnée ». Ainsi, l'initiative Data for Common Purpose (World Economic Forum, 2021) vise à produire un cadre de gouvernance pour améliorer les avantages sociétaux des données. C'est aussi le cas de la loi européenne sur la gouvernance des données (European Commission, 2020) ou encore le plan d'action de l'US Federal Data Strategy (Federal Data Strategy, 2022) qui vise à mieux exploiter les données en tant qu'actif stratégique.

Plusieurs méthodes d'évaluation de la valeur des données ont été proposées (Moody et Walsh, 1999 ; Ciuriak, 2019 ; Wang et Zhao, 2020 ; Otto, 2015). Une première famille est composée de techniques d'estimation de la valeur au travers des coûts que la gestion des données requiert (coût de production et de stockage des données, coût de remplacement des données obsolètes, etc.) et de leur impact sur le flux de trésorerie. L'un des principaux avantages de cette famille de méthodes est sa facilité

d'utilisation. La deuxième famille cible l'évaluation de la valeur marchande des données, fondée sur l'estimation du prix que les entreprises paient pour des données comparables sur le marché. Simples à mettre en œuvre, ces méthodes ne s'appliquent pas à toutes les données dans la mesure où certaines données ne sont tout simplement pas échangeables, parce qu'elles représentent pour les entreprises un avantage concurrentiel. De plus, pour obtenir un prix réel des données, il faut qu'il existe un marché efficace, ce qui n'est pas toujours le cas. En outre, il convient de rappeler que le prix n'est pas synonyme de valeur. La troisième famille est caractérisée par l'approche de la valeur économique (Garifova, 2015) qui peut être utilisée pour identifier la valeur ajoutée de la donnée, par exemple à des fins commerciales ou pour des cas d'usage spécifiques. Comme pour les valeurs obtenues via les autres approches, une grande partie de cette valeur est subjective.

Les approches précédentes appliquent les techniques classiques d'évaluation des actifs tangibles à la donnée, qui est par nature intangible. D'autres approches, fondées sur les externalités, proposent de considérer la donnée à un niveau plus large, par exemple en considérant les bénéfices tirés par la société de l'utilisation de ces données (Coyle et Diepeveen, 2021 ; Antuca & Noble, 2021). C'est le cas notamment des données ouvertes. A notre connaissance, aucune approche d'évaluation de la valeur de la donnée proprement dite ne s'appuie explicitement sur les risques associés. A noter toutefois qu'une étude comparative de ces approches y insère aussi les approches d'évaluation du risque (Bodendorf *et al.*, 2022). Ainsi, il n'existe pas d'approche holistique qui combine tous ces points de vue pour produire une évaluation de la valeur de la donnée. Cette combinaison suppose toutefois que les trois dimensions soient ramenées à une même unité qui est financière. C'est précisément cette évaluation holistique que nous proposons de faire dans cet article.

### **3. Notre approche**

Nous proposons ci-dessous une approche facilitant le processus d'évaluation de la valeur des données. Elle s'appuie sur trois artefacts. Le premier est un nouveau construit représentant la valeur de la donnée. Le second consiste en un modèle conceptuel regroupant l'ensemble des informations qui interviennent dans le processus d'évaluation de la donnée. Le troisième est la méthode proprement dite qui est fondée sur le modèle conceptuel des données et structurée suivant l'approche hiérarchique multicritère. La validation de notre approche est décrite ensuite.

#### **3.1. La valeur de la donnée**

La donnée, actif de l'entreprise, nécessite la mise en place d'une gouvernance au même titre que les ressources humaines, le patrimoine immobilier ou tout autre actif stratégique. La gouvernance de la donnée peut ainsi être définie par un triple objectif : maximiser sa valeur en minimisant les risques et les coûts associés (Tallon, 2013). La valeur est ainsi définie selon trois dimensions : juridique (obligation de conformité et gestion des risques), métier (pertinence pour le fonctionnement et les revenus de l'entreprise) et historique (coûts et impact sur les flux de trésorerie) alors

que les méthodes d'évaluation de la littérature n'intègrent qu'une seule de ces dimensions : la valeur métier ou la valeur historique fondée sur les coûts. Les méthodes d'évaluation des risques sont cantonnées dans le domaine de la sécurité des systèmes d'information mais ne sont jamais agrégées aux approches d'évaluation de la valeur. C'est pour y remédier que nous proposons de définir la *valeur de la donnée* comme un concept à trois dimensions :

- la *valeur* produite par la donnée pour l'organisation qu'il s'agisse de sa valeur d'usage (l'utilisation de la donnée produit une valeur) ou de sa valeur d'échange (la donnée peut être vendue en tant que telle ou associée à un produit ou un service) ;
- le *coût* engendré par la donnée qui ne se limite pas au coût d'acquisition mais inclut tous les aspects de son cycle de vie (acquisition, production, stockage ou archivage, utilisation, partage, destruction). Ce coût vient en diminution de la valeur produite par la donnée ;
- enfin, le *risque* lié à la donnée. Il est légal ou réglementaire (un non-respect des lois ou réglementations liées aux données peut générer une perte financière). Il peut aussi être stratégique (une mauvaise décision prise à cause de données erronées) ou entacher la réputation de l'organisation (un site web de mauvaise qualité impacte négativement l'image). Enfin, les risques opérationnels incluent des aspects liés aux données : de nombreuses activités des entreprises sont dépendantes de la disponibilité et de la qualité des données qu'elles utilisent.

Pour faciliter le processus d'évaluation de la valeur des données, nous présentons ci-dessous un modèle conceptuel qui s'appuie ces trois dimensions de la valeur.

### 3.2. Le modèle conceptuel de données

Le modèle conceptuel a pour objectif de décrire et relier toutes les informations nécessaires à l'évaluation de la valeur de la donnée (Figure 1). Il décrit donc les trois dimensions de la valeur présentées précédemment. La *donnée* elle-même est décrite. On distingue la donnée interne de la donnée externe, dont les méthodes d'évaluation sont différentes. D'autres attributs qui interviennent dans le processus d'évaluation sont représentés tels que le volume, l'accessibilité, la fréquence d'usage et l'exclusivité. La donnée est relative à un domaine (par exemple le domaine client, le domaine RH, etc.). Elle est caractérisée par des facteurs de qualité (fraîcheur, précision, etc.) qui impactent sa valeur. Sa nature (donnée numérique, image, donnée capteur, etc.) est aussi un élément important.

Le deuxième concept est celui de *dimension*. Une donnée est évaluée selon trois dimensions : le coût, la valeur et le risque. Sa valeur nette résulte de l'agrégation de ces trois dimensions. La valeur d'usage représente la capacité de l'entreprise à améliorer son efficacité interne ainsi que le potentiel de développement du business qu'elle peut générer ou encore la capacité de croissance externe (alliances, fusions,

etc.) qui en résulte. La valeur d'échange traduit la possibilité de vendre, de louer voire de mettre à disposition gratuitement ses données. L'évaluation du risque et du coût sont nécessaires pour le calcul de la valeur nette.

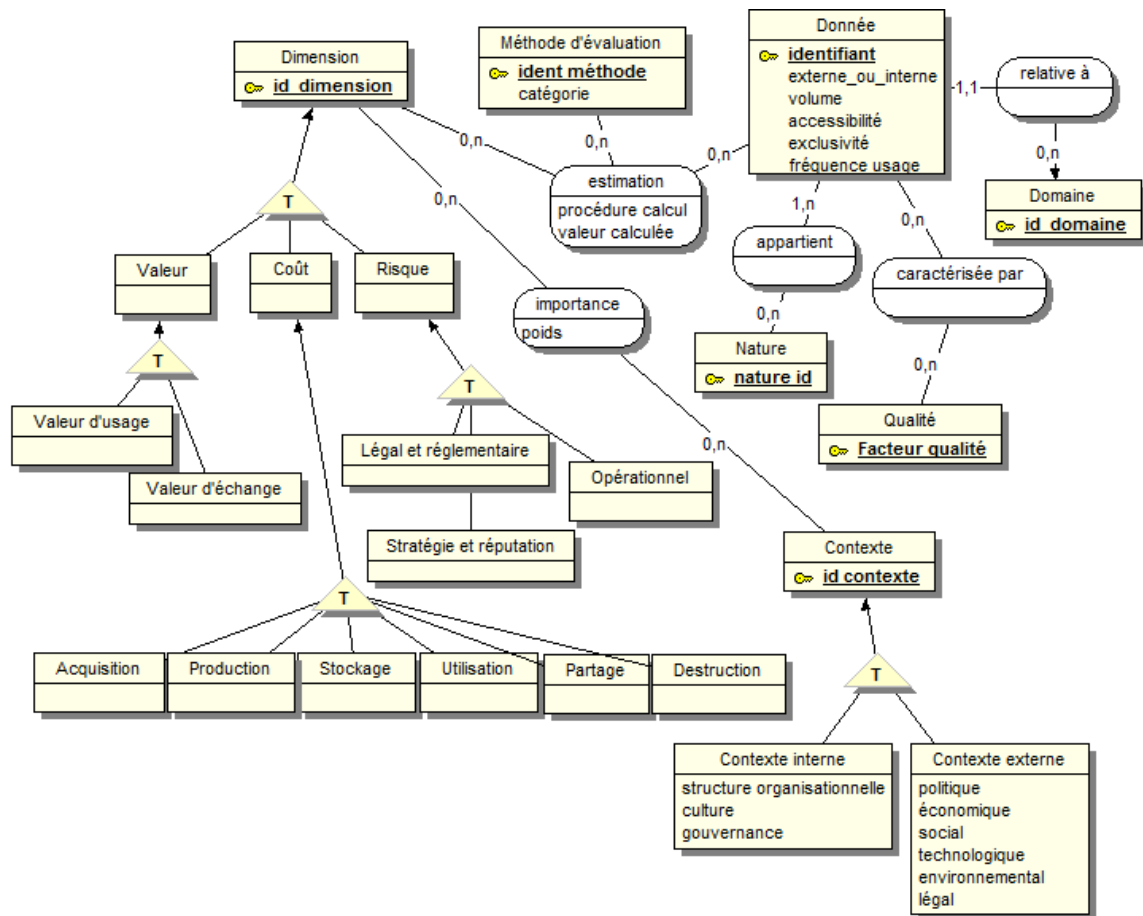


Figure 1. Le modèle conceptuel de données

Le troisième concept est celui du *contexte*. Ce dernier joue un rôle important dans le processus d'évaluation de la valeur des données. Il existe un lien et une certaine interdépendance entre le contexte et les processus d'évaluation. Le contexte peut être interne à l'entreprise qui évalue les données. La structure organisationnelle, la culture et la gouvernance de l'entreprise jouent alors un rôle important dans le processus d'évaluation. Ainsi, un organisme public n'obéit pas aux mêmes règles

dans l'évaluation de ses actifs qu'une entreprise privée. Le contexte externe joue un rôle tout aussi important. Nous le caractérisons par les attributs de type PESTEL (Politique, Economique, Social, Technologique, Environnemental et Légal) (Aguilar, 1967). Le contexte détermine la pertinence et l'importance relatives des différentes dimensions de l'évaluation. Ainsi, dans certains contextes, par exemple un établissement public, la vente de données n'est pas envisageable. Dans le cadre bancaire, les aspects liés aux risques de la donnée sont plus prégnants que dans le secteur de la distribution par exemple. C'est la raison pour laquelle le modèle contient une relation binaire entre la dimension et le contexte caractérisée par un poids qui peut être nul quand la dimension n'est pas mobilisable dans le contexte et qui permet, le cas échéant, de combiner différentes dimensions pour une évaluation globale. Ainsi ces poids permettent d'annuler l'impact d'un critère ou sous-critère ou de le nuancer quand, par exemple, un risque est moindre qu'un autre dans un contexte donné.

L'*estimation* de la valeur est une relation ternaire entre la donnée, la dimension et la *méthode d'évaluation*. Ses attributs principaux sont la procédure de calcul et la valeur calculée. Comme nous l'avons remarqué dans l'état de l'art, il existe un nombre important de méthodes d'évaluation de la donnée. Les méthodes d'évaluation appartiennent à des catégories complémentaires. Cinq catégories sont mobilisées dans notre processus d'évaluation : les approches fondées sur le revenu, celles fondées sur les coûts, les approches de type marché, les méthodes d'analyse de risque et les approches par externalités (Coyle et Diepeveen, 2021).

Une des contributions principales de ce modèle, au-delà de la prise en compte du risque combiné avec le coût et la valeur, est l'intégration du contexte dans le processus d'évaluation de la valeur.

Fondé sur ce modèle conceptuel, nous présentons dans le paragraphe suivant l'arborescence multicritères qui guide le processus d'évaluation. Elle permet au décideur une vision plus holistique de l'évaluation de la valeur et la possibilité de tester voire de combiner plusieurs scénarii.

### **3.3. La méthode d'évaluation**

L'évaluation de la valeur des données permet aux entreprises d'avoir une vue d'ensemble de leur patrimoine informationnel. Elle facilite la comparaison de la valeur des données (actifs intangibles) avec celle des actifs tangibles. Elle permet également de comparer la valeur des données pour différents cas d'usage. Nous avons choisi de représenter le modèle d'aide à la décision sous la forme d'une hiérarchie de critères (Saaty, 1994). Les raisons principales de ce choix résident dans la capacité de la démarche à structurer de manière hiérarchique le problème d'évaluation de la valeur de la donnée, qui est complexe du fait des nombreux critères et sous critères qui interviennent dans cette évaluation. De plus, la démarche permet une comparaison binaire des différentes alternatives en structurant les

priorités représentées par des poids. Enfin, la démarche permet une analyse de sensibilité plus aisée. En effet, les critères et les sous-critères peuvent avoir des poids variables et leur nombre n'est pas limité. Il est donc possible au décideur de modifier la valeur d'un critère ou d'ajouter ou de supprimer des critères. Pour des raisons d'espace nous ne présentons que les deux premiers niveaux de la hiérarchie (Fig. 2). Les autres niveaux sont détaillés dans les tableaux 1, 2 et 3.

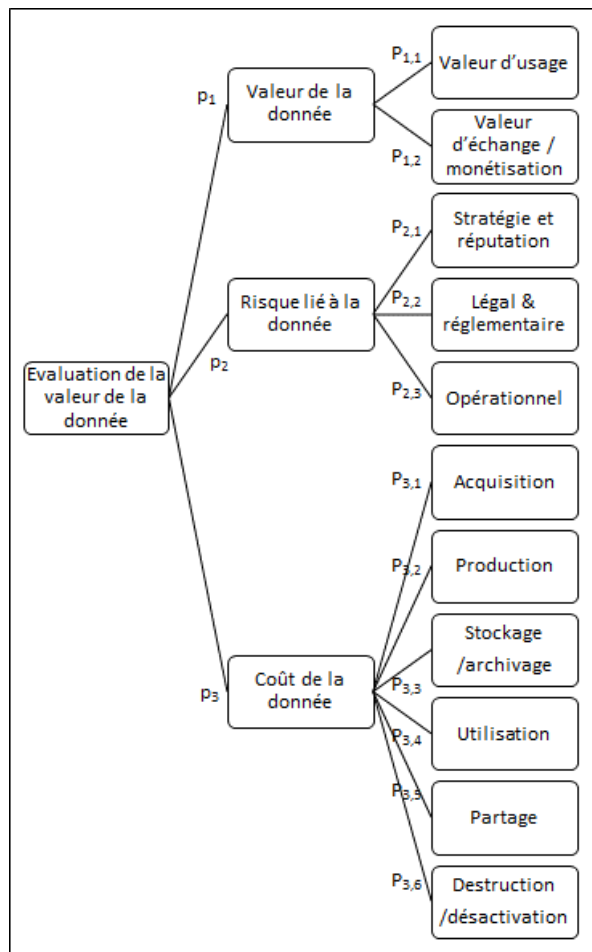


Figure 2. La hiérarchie de critères

Chaque nœud terminal de la hiérarchie de la figure 2 est lui-même décomposé en sous-critères. Puis plusieurs scénarii sont proposés pour faciliter l'évaluation de ces critères. Enfin, un exemple de métrique est associé à chaque scénario.

Tableau 1. Décomposition de la hiérarchie (dimension Valeur)

Sous-critère	Décomposition du sous-critère	Exemple de scénario d'illustration	Exemple de métrique	
Valeur d'usage (nb d'utilisations, utilisations prévues, utilisations effectives, etc.)	améliorer l'efficacité interne (différentiel d'efficacité)	améliorer les processus	temps gagné dans la livraison d'un produit	
		optimiser les tâches	temps gagné dans l'exécution de la tâche	
		réduire les coûts fournisseurs	augmentation du ratio de "gros" fournisseurs	
		réduire les coûts de production	augmentation juste-à-temps	
		optimiser les RH	diminution du délai de recrutement	
	se développer par croissance interne (augmentation parts de marché)	pénétrer le marché	augmentation chiffre d'affaires produits	
		étendre le marché	chiffre d'affaires nouveaux marchés	
		étendre les produits/services	chiffre d'affaires nouveaux produits/services	
		diversifier les produits/services	vente d'études fondées sur les données	
	développer par croissance externe	nouer des alliances	part de marché additionnelle	
		procéder à des fusions/acquisitions	économie d'échelle	
	Valeur d'échange / monétisation (chiffre d'affaires)	vendre/louer des données (chiffre d'affaires)	vendre données de consommation	chiffre d'affaires réalisé
			vendre données métier	
vendre données de déplacement				
vendre données de navigation Internet				
donner des données (externalité)		mettre à disposition données ouvertes	amélioration du bien commun	
		mettre à disposition données ESG		

Les tableaux 1, 2 et 3 décrivent successivement les branches Valeur, Risque et Coût ainsi décomposées. La première colonne de chaque tableau reprend les sous-critères déjà décrits. Ainsi, par exemple, la dimension Valeur (Tableau 1) se décline en deux familles liées respectivement à l'utilisation de la donnée (valeur d'usage) et sa monétisation (valeur d'échange). La seconde colonne décompose chaque famille à un niveau plus fin. Ainsi, l'usage de la donnée peut améliorer l'efficacité interne et/ou permettre la croissance interne de l'entreprise par développement de ses produits/services et marchés et/ou encore contribuer à sa croissance externe.

Le tableau 2 décrit la dimension Risque qui se décline en trois familles liées respectivement à l'impact négatif de la donnée sur la stratégie et la réputation, à l'aspect légal et réglementaire et en termes de risques opérationnels.

Tableau 2. Décomposition de la hiérarchie (dimension Risque)

Sous-critère	Décomposition du sous-critère	Exemple de scénario d'illustration	Exemple de métrique
Stratégie et réputation	ESG (Environnement-Social-Gouvernance)	augmenter l'empreinte carbone	variation du taux de respect des engagements carbone
		faire de la discrimination via les données	mesure du risque psychosocial par enquête
		diffuser des données erronées sur la gouvernance	variation de l'indice de satisfaction des actionnaires
	Ethique	diffuser des informations trompeuses	mesure de l'évolution de l'image de l'entreprise
		informer de manière non objective sur les produits	mesure de l'évolution de la perception des produits
	Confiance	diffuser des informations erronées	variation de l'indice de confiance client
		subir des fuites d'informations sensibles	mesure de l'évolution de l'image de l'entreprise
	Prise de décision	décider avec des données erronées	différentiel de performance globale
		décider avec des données obsolètes	
		prendre des décisions en l'absence de données	
Légal & réglementaire	Conformité à la loi	diffuser des informations discriminatoires	mesure financière du risque juridique encouru
		diffuser des informations confidentielles	
	Conformité aux réglementations	enfreindre le RGPD	mesure de la conséquence financière
		enfreindre le Protection of Personal Information Act (APPI)	
Propriété intellectuelle	brevet	mesure de la conséquence financière	
	enfreindre les droits d'auteur		
Opérationnel	Personnes	commettre des malveillances dans les données	évaluation des pertes attendues et des pertes exceptionnelles
		commettre des erreurs dans les données	
	Processus	ralentir les processus à cause de données manquantes	
		perturber les processus à cause de données erronées	
	Systèmes et technologies	diffuser des informations non autorisées	
interrompre des systèmes à cause de données manquantes			

A titre d'exemple, l'aspect légal et réglementaire est lui-même décomposé en la conformité aux lois, à la réglementation et à la propriété intellectuelle. Chaque composant est lui-même illustré par plusieurs scénarii. Ainsi la conformité aux



réglementations est illustrée par les scénarii « enfreindre le Règlement Général sur la Protection des Données (RGPD) » et « enfreindre le règlement japonais Protection of Personal Information Act (APPI) ».

Tableau 3. Décomposition de la hiérarchie (dimension Coût)

Sous-critère	Décomposition du sous-critère	Exemple de scénario d'illustration	Exemple de métrique
Acquisition	Données externes	payer les données	prix d'achat
		collecter des données (enquête)	coût de l'enquête externe
	Données internes	saisir des données	nombre d'heures*coût heure saisie
		collecter des données (enquête interne)	coût de l'enquête interne
Production	Données opérationnelles	contrôler les données	effort de vérification
		corriger les données erronées	effort de correction
	Données décisionnelles	nettoyer les données pour l'analytique	effort de nettoyage
		transférer les données de l'opérationnel vers le décisionnel	coût de maintenance des processus de flux
Stockage/archivage	Stockage des données	stocker les données sur site	coût total * part volume
		stocker les données sur le cloud	montant contrat
	Sécurisation des données	sauvegarder les données	coût des processus de sauvegarde
		anonymiser les données sensibles	effort d'anonymisation
Utilisation	Données opérationnelles	traiter les données opérationnelles	quote-part des frais IT
		vérifier les données traitées	effort de vérification
	Données décisionnelles	visualiser les données	coût des processus de visualisation
		calculer les indicateurs	quote-part des efforts de business analytics
Partage	Acheminement par courrier	acheminer l'information par voie postale	frais d'affranchissement
		acheminer l'information par email	quote-part coût serveur de messagerie
	Mise à disposition	virtualiser les données	coût architecture de virtualisation
		maintenir l'architecture de partage des données	coût architecture de données
Sites web et plateformes électroniques	mettre à jour les contenus électroniques	coût humain et technique gestion des contenus	
	mettre à disposition l'information sous forme accessible	coût de mise en conformité et accessibilité	
Destruction/désactivation	Données sensibles	écraser les données	coût destruction "forte"
	Données non sensibles	effacer les données	coût standard destruction

Puis, la perte de confiance liée à un incident relatif aux données se traduit par une métrique de variation de l'indice de confiance client. L'importance du risque n'est

pas nécessairement la même selon le domaine d'activité de l'organisation. C'est pourquoi toutes ces mesures doivent être pondérées en fonction du contexte, comme expliqué plus loin.

Enfin, le tableau 3 détaille la dimension Coût de la donnée selon son cycle de vie. La colonne suivante décrit les caractérisations de la donnée permettant ensuite de décliner les différents scénarii, composants du coût, puis d'associer des métriques (dernière colonne). Ainsi, la destruction des données se décline en scénarii qui diffèrent en fonction de la sensibilité des données (effacer les données ou écraser les données). Les métriques de coût associées (coût destruction forte et coût standard) sont spécifiques.

Une des originalités de l'approche est de combiner cinq familles de méthodes d'évaluation de la valeur de la donnée. Ces familles sont matérialisées par la nuance de couleur de la deuxième colonne des trois tableaux : blanc pour les approches fondées sur le revenu, gris clair pour celles fondées sur les coûts, tacheté pour les approches de type marché, gris foncé pour les méthodes d'analyse de risque et hachuré pour les approches par externalités. Même si les approches utilisées proviennent de familles différentes, toutes les métriques sont financières, ce qui permet de les comparer, de les combiner avec des pondérations choisies et de les agréger pour construire un indicateur composite de la valeur.

La procédure d'évaluation (Saaty 1994) comprend donc les phases suivantes :

- Pour chaque nœud de la hiérarchie, à l'exclusion de la racine, le contexte détermine l'importance de chaque critère.
- Nous procédons ensuite à l'évaluation proprement dite de chaque critère se trouvant sur les nœuds terminaux de la hiérarchie.
- Puis nous évaluons la valeur de tous les autres nœuds inclus la racine de l'arborescence en procédant au calcul de la somme pondérée.

A noter qu'à certains nœuds de l'arborescence, peuvent être associées plusieurs métriques. Il peut s'agir ainsi de mesures résultant de l'agrégation des mesures des niveaux inférieurs de l'arborescence ou de mesures alternatives. Dans ce cas, il faut soit choisir une des mesures, soit les combiner. A titre d'exemple, la valeur d'usage des données peut être mesurée par le nombre d'utilisations effectives de celles-ci ou par l'évaluation de l'efficacité accrue de l'organisation grâce à ces utilisations.

### **3.4. La validation**

La contribution de cet article consiste en une approche fondée sur trois artefacts : un construit (concept de valeur de la donnée), un modèle conceptuel et une méthode d'évaluation de la valeur. Nous proposons de valider l'approche par quelques cas d'usage décrits ci-après.

Une chaîne de salons d'esthétique a acheté des données décrivant les comportements d'achat des consommateurs dans le secteur de la beauté (Pwc, 2019). En étudiant ces

données, l'entreprise a pu corréler l'âge des clients et le type de campagnes marketing auxquelles ceux-ci sont les plus réceptifs. Ainsi, elle a recruté des influenceurs sur les réseaux sociaux pour atteindre les femmes de moins de 30 ans et faire la promotion des produits cosmétiques et de maquillage. Pour cibler les femmes de plus de 50 ans, elle a fait appel à des actrices célèbres et mis en avant des traitements de raffermissement de la peau. Par cette segmentation, l'entreprise a augmenté de 20% son chiffre d'affaires annuel. Cette évaluation de la valeur correspond à la mesure du différentiel en parts de marché obtenu par une meilleure pénétration de celui-ci (composant *pénétrer le marché* de l'action de *croissance interne* via l'utilisation de données – *valeur d'usage*). L'utilisation de notre approche consiste à enrichir cette dimension Valeur en intégrant d'une part les risques potentiels (par exemple, l'utilisation de données non conformes au RGPD) et d'autre part les coûts associés (acquisition des données, production des informations incluant la préparation, analyse, destruction, etc.). Cet exemple valide ainsi partiellement notre approche et en montre aussi la valeur ajoutée par la prise en compte d'une évaluation holistique.

Une société californienne de tests ADN a vendu un accès exclusif à sa base de données, comprenant les génomes de plus de 5 millions de personnes à un géant pharmaceutique pour 300 millions de dollars (PwC, 2019). La base de données peut être utilisée pour la recherche et le développement de nouveaux médicaments. Grâce à cette transaction, on peut ainsi estimer qu'une base de données génétiques comparable peut être vendue à 60 dollars par ligne de données. Cet exemple corrobore la métrique utilisant le chiffre d'affaires réalisé grâce à la vente d'un jeu de données (composant *vendre des données métier* de l'action de *vendre/louer des données* via la monétisation des données – *valeur d'échange*). Toutefois, notre approche permet de mettre l'accent sur le risque important tant en conformité à la réglementation RGPD qu'à l'éthique ou la confiance qui pourraient être compromises par cette vente si l'anonymisation complète des données n'était pas garantie. Cette anonymisation est particulièrement coûteuse quand elle concerne des données génétiques quasi-identifiantes. Ainsi notre approche est validée en ce qui concerne la monétisation et l'exemple montre sa contribution supplémentaire dans l'évaluation des coûts et des risques induits.

Ces deux cas illustrent l'approche et constituent les premiers pas d'une validation de son utilité. L'arborescence guide l'évaluation en parcourant depuis la feuille jusqu'à la racine. Puis le parcours de haut en bas permet d'amplifier le raisonnement de l'évaluateur en sollicitant les branches complémentaires pour une évaluation plus holistique. D'autre part, ils nous ont permis de valider la complétude du modèle conceptuel et, par-là, son utilité.

La validation de notre approche auprès d'experts nous a permis à la fois de souligner l'intérêt de combiner l'approche par les risques et celles par les coûts et la valeur mais aussi de mettre le doigt sur la difficulté d'assurer une cohérence dans l'évaluation quand celle-ci est effectuée par des parties prenantes de culture différente. Ainsi, l'appréhension du risque est différente par les juristes et par les

experts du marketing. Le cabinet de conseil que nous avons interrogé est surtout intéressé par l'approche qui structure sa pensée et lui permet d'améliorer sa proposition de valeur pour des clients ou prospects. Le directeur des données d'une grande entreprise d'assurance que nous avons aussi confronté à l'approche a, quant à lui, vu un moyen convaincant de présenter à ses dirigeants ses demandes de budgets tant pour valoriser les données que pour gérer les risques associés.

#### 4. Conclusions et recherches futures

Les principales contributions de cet article sont :

- la définition d'un concept de valeur de la donnée intégrant les trois dimensions du risque, du coût et de la valeur,
- la proposition d'un modèle conceptuel qui relie l'ensemble des concepts participant au processus d'évaluation de la valeur des données, et notamment le contexte interne et externe de l'organisation concernée,
- le développement d'une arborescence hiérarchique multicritère qui sert de base à un système d'aide à l'évaluation holistique de la valeur des données d'une organisation,
- la validation de l'approche à l'aide de cas d'usage fondés sur des cas réels décrits dans la presse professionnelle. Cela ne constitue qu'une première étape dans la validation de l'approche proposée.

Il est bien entendu qu'on ne peut prétendre à l'exhaustivité dans ce type d'approche d'évaluation d'un concept très évolutif. Toutefois, le modèle proposé est extensible sans contrainte. L'introduction d'une nouvelle dimension, d'un nouveau critère ou sous-critère n'impacte que l'éventuelle adaptation des poids associés aux autres nœuds du même niveau de l'arborescence. Cet arbre pondéré permet aussi de garantir l'adaptabilité de l'approche à toutes les organisations, publiques ou privées, quel que soit le secteur, au prix de la redéfinition des poids.

Nous prévoyons notamment des ateliers pour compléter les scénarii et l'application de la démarche à plusieurs exemples réels et le développement d'un outil d'automatisation de la démarche facilitant le processus de décision des parties prenantes de l'entreprise. Ces étapes permettront une validation conjointe et un raffinement des différents artefacts de l'approche (concept, modèle, méthode). Un autre axe de recherche future sera d'affiner la description des données (domaine, nature, qualité) en s'appuyant, le cas échéant, sur les normes existantes. Cette approche ainsi confortée pourra utilement alimenter l'évaluation des entreprises en cas de rachat et contribuer à la base de connaissances sur les stratégies de valorisation des actifs informationnels des entreprises.

**Remerciements.** Les auteurs remercient les partenaires de la Chaire Stratégie et Gouvernance de l'Information de l'ESSEC, au sein de laquelle cette recherche a été réalisée.

## Bibliographie

Aguilar, F. J. (1967). *Scanning the business environment*. Macmillan.

Antuca, A., Noble, R. (2021). Data: how it affects competitive dynamics, how to value it, and whether to provide third-party access to it. *Competition Law Journal*, 20(2), 102-110.

Attard, J., Brennan, R. (2018). A Semantic Data Value Vocabulary Supporting Data Value Assessment and Measurement Integration. *Proceedings of the 20th International Conference on Enterprise Information Systems (ICEIS)*, pages 133-144.

Bodendorf, F., Dehmel, K., Franke, J (2022). Scientific Approaches and Methodology to Determine the Value of Data as an Asset and Use Case in the Automotive Industry, *Proceedings of the 55th Hawaii International Conference on System Sciences*.

Ciuriak, D. (2019). Unpacking the Valuation of Data in the Data-Driven Economy, *Conference on Global Data Law*, New York, 26-27.

Corrado, C. (2019) Data as an Asset: Expanding the Intangible Framework, *Conference on the Economics, Governance and Management of AI, Robots and Digital Transformation (EMAAE)*.

Coyle, D., Diepeveen, S. (2021). "Creating and governing social value from data." Available at SSRN 3973034.

European Commission (2020). *Data Governance Act, Proposal for a Regulation of the European Parliament and of the Council on European data governance (Data Governance Act) — COM/2020/767 final*. Brussels, Belgium: European.

Federal Data Strategy (2022). *Data, accountability, and transparency: creating a data strategy and infrastructure for the future*, <https://strategy.data.gov>, accédé le 26.02.2022.

Garifova, L.F. (2015). Infonomics and The Value of Information in The Digital Economy, *Procedia Economics and Finance* 23 (pp. 738 – 743).

Moody, D., Walsh, P. (1999). Measuring The Value Of Information: An Asset Valuation Approach, *European Conference on Information Systems*, ECIS'99, pp.1-17.

Otto, B. (2015). Quality and Value of the Data Resource in Large Enterprises, *Information Systems Management*, 32 (pp. 234–251).

PwC (2019). *Putting a value on data*, <https://www.pwc.co.uk/data-analytics/documents/putting-value-on-data.pdf>.

Saaty, T. L. (1994). How to make a decision: the analytic hierarchy process. *Interfaces*, 24(6), (pp. 19-43).

Savona, M. (2019). *The Value of Data: Towards a Framework to Redistribute It*, SPRU Working Paper Series SWPS 2019-21 (Octobre 2019).

Short, J., Todd, S. (2017). What's your data worth? *MIT Sloan Management Review*, Spring.

Tallon, P. P. (2013). Corporate governance of big data: Perspectives on value, risk, and cost. *Computer*, 46(6), (pp. 32-38).

Wang, Y., Zhao, H. (2020). Data Asset Value Assessment Literature Review and Prospect, *Journal of Physics: Conference Series*, 1550.

Wdowin, J., Diepeveen, S. (2020). *The Value of Data – Literature review*, Bennett Institute for Public Policy, Cambridge.

World Economic Forum (2021). *Data for Common Purpose: Leveraging Consent to Build Trust*, White paper, Novembre 2021.

Zeiter, A., Hagel, J., Snyder, S (2021). *Articulating Value from Data*, White Paper, World Economic Forum, Novembre 2021.

---

# Application de l'Ingénierie des Exigences basée sur les Modèles dans Trois Grands Projets Collaboratifs Européens : Un Rapport d'Expérience

Andrey Sadovykh<sup>1</sup>, Hugo Bruneliere<sup>2</sup>, Dragos Truscan<sup>3</sup>

1. SOFTEAM

Paris, France

andrey.sadovykh@softeam.fr

2. IMT Atlantique & LS2N (UMR CNRS 6004)

Nantes, France

hugo.bruneliere@imt-atlantique.fr

3. Åbo Akademi University

Turku, Finland

dragos.truscan@abo.fi

---

REFERENCE DE L'ARTICLE INTERNATIONAL. Cet article est un résumé de l'article :

Andrey Sadovykh, Dragos Truscan, Hugo Bruneliere:

*Applying Model-based Requirements Engineering in Three Large European Collaborative Projects: An Experience Report*. RE 2021: 367-377

MOTS-CLES : Ingénierie des Exigences, Ingénierie basée sur les Modèles, Projets Collaboratifs, Rapport d'Expérience, Passage à l'Echelle, Hétérogénéité, Traçabilité, Automatisation

---

Ce papier rapporte notre expérience pratique de proposition et d'application d'une approche et d'un langage d'Ingénierie des Exigences basées sur les Modèles. La période concernée de 5 ans couvre trois grands projets collaboratifs européens, chacun d'entre eux fournissant diverses solutions logicielles complexes (e.g., frameworks, ensemble d'outils intégrés, etc.). Un élément clé des projets européens est l'hétérogénéité de leurs participants. Ces derniers proviennent de différents domaines d'applications, ont différentes tailles, divers niveaux de maturité ou types d'expertise de recherche. Malgré tout, ils doivent travailler ensemble afin de réaliser un ensemble d'objectifs de R&D, habituellement validés par plusieurs cas d'étude utilisés comme plateforme commune pour expérimenter sur des technologies nouvellement conçues et développées. Cependant, la diversité et le nombre de ces

partenaires impliquent aussi des défis en matière de gestion de projet liés à 1) l'élicitation des besoins provenant des fournisseurs de cas d'utilisation industriel et 2) l'identification non seulement de solutions concrètes à fournir pendant le projet (durant typiquement 3 ans) mais aussi d'une feuille de route pour le développement de la solution technique finale. Lorsque de tels défis ne sont pas traités correctement, ils peuvent influencer négativement les résultats du projet.

Afin de répondre à ces défis, nous avons proposé une approche reposant sur l'utilisation d'un langage de modélisation dédié à l'Ingénierie des Exigences pour de tels projets. Les solutions techniques sont ainsi modélisées sous la forme d'un *Framework* ou ensemble d'outils dont les exigences sont élicitées à partir de celles de cas d'étude industriels. Le framework agrège des *Tool Components* (Composants Outil) applicables dans un ou plusieurs cas d'étude et décrits via des *Tool Component Requirements* (Exigences de Composant Outil). Un composant outil a une architecture incluant les interfaces pour son interconnexion avec d'autres outils, ses priorités de développement et différentes échéances décrites dans sa *Tool Component Roadmap* (Feuille de Route du Composant Outil). Grâce à cette modélisation, l'équipe de coordination technique peut ainsi concevoir la *Framework Architecture* (Architecture du Framework) et la *Framework Development Roadmap* (Feuille de Route de Développement du Framework) permettant ensuite aux fournisseurs de cas d'étude de créer des *Case Study Requirements Validation Roadmaps* (Feuilles de Route de Validation des Exigences d'un Cas d'Etude). Lorsque des *Tool Component Implementations* (Implémentations de Composant Outil) deviennent disponibles, elles sont finalement intégrées et évaluées au regard des cas d'étude via un processus dédié.

L'approche et le langage de modélisation proposés ont été implémentés sur l'environnement de modélisation Modelio, puis mis en œuvre de manière effective dans les trois projets collaboratifs européens DataBio, REVaMP2 et MegaM@Rt2. Nous les avons ensuite évalués via un sondage auprès des participants à ces trois projets, ainsi que par l'analyse des données de modélisation collectées à la fin de ces projets. En conclusion, l'utilisation de notre approche a montré des bénéfices significatifs en matière de gestion de projet ou encore de définition de l'architecture de la solution globale. Le passage à l'échelle dans le contexte de gros projets, le meilleur support de l'hétérogénéité de leurs participants ainsi que des capacités intéressantes de traçabilité de l'information modélisée et d'automatisation (notamment pour la génération semi-automatisée de livrables de projet) ont aussi été mis en avant lors de ce travail.

#### *Remerciements*

*Ce travail a été financé par l'ECSEL JU (projet MegaM@Rt2, No. 737494) (projet AIDOaRt, No. 101007350), par ITEA3 (projet REVaMP2, No. 15010) et par H2020 (projet DataBio, No. 732064) (projet VeriDevOps, No. 957212).*



---

# Perception des Méthodes Agiles par les Développeurs Aujourd'hui

**Florian Gauthier, Rébecca Deneckère**

*Centre de Recherche en Informatique (CRI),  
Université Paris 1 - Panthéon-Sorbonne  
12 Place de Panthéon, 75005, Paris, France  
florian.gauthier.learning@outlook.fr, rebecca.deneckere@univ-paris1.fr*

---

*RESUME. Les méthodes agiles sont extrêmement répandues mais leurs avantages et limites ne sont pas toujours bien connus. Les développeurs sont amenés à travailler quotidiennement avec les méthodes agiles au cours de leurs projets et il est intéressant de comprendre comment ils perçoivent cette organisation de projet flexible. L'objectif de ce travail est donc d'étudier la perception des méthodes agiles par cette population en particulier. Nous avons créé et lancé un questionnaire pour tenter d'identifier la perception qu'avaient les développeurs sur l'utilisation de ces méthodes et la comparer à d'autres études faites sur le sujet dans la littérature scientifique.*

*ABSTRACT. Agile methods are present everywhere now but their advantages and limitations are not so well known. The developers work everyday with agile methods in their projects and it is interesting to understand how they apprehend this flexible organization. The goal of this work is then to study this perception of agile methods by this specific population. We created and used a survey to identify the perception of developers on the use of agile methods. We compared the results to other studies in the scientific literature.*

*Mots-clés : Questionnaire ; Développeur ; Méthode Agile*

*KEYWORDS: Survey ; Developer ; Agile Method*

---

## 1. Introduction

De tout temps, les entreprises ont cherché des moyens de mener à bien leurs projets pour accomplir leurs objectifs. Pour ce faire, elles ont appris à organiser le travail au travers de la gestion de projet. Il existe différentes méthodes de gestion de projets comme le modèle en cascade ou le cycle en V qui sont des méthodes parmi les plus populaires pouvant être utilisées dans des projets informatiques. Cependant, ces méthodes dites traditionnelles ne conviennent pas forcément à tous les projets et sont critiquées du fait de leur rigidité et du manque de visibilité de l'avancement du projet (Augustine *et al.*,

2005). C'est alors qu'émergèrent plusieurs méthodes plus flexibles et adaptées au développement logiciel, qualifiées plus tard de méthodes agiles par le manifeste agile (Beck *et al.*, 2001) en 2001. Parmi les méthodes agiles les plus populaires on retrouve Scrum (Schwaber, 1995), EXtreme Programming (Beck, 2000) et Kanban (Anderson, 2003), mais d'autres méthodes existent, certains projets mettant également parfois en œuvre uniquement certains principes agiles spécifiques sans utiliser l'ensemble des principes préconisés par le manifeste.

Il existe de nombreux travaux de recherche menés sur le sujet des méthodes agiles (cf. section 2 sur les travaux connexes). On trouve des études sur leurs forces et faiblesses, sur leur acceptation en entreprise, sur les facteurs de succès et d'échec des projets agiles, sur le développement agile à grande échelle ou sur la perception qu'ont les personnes de ces méthodes. Nous ne trouvons cependant pas de travaux sur le ressenti spécifique des développeurs et c'est ce dernier aspect qui nous intéresse ici. Les développeurs sont amenés à travailler quotidiennement avec les méthodes agiles au cours de leurs projets et il est intéressant de comprendre comment ils perçoivent cette organisation de projet flexible. L'objectif de ce travail est donc d'étudier la perception des méthodes agiles, quelles qu'elles soient, par cette population en particulier. Le travail s'articulera autour de la problématique suivante : **Comment sont perçues les méthodes agiles par les développeurs aujourd'hui ?** Nous parlerons des travaux connexes dans la section 2 et présenterons la méthode de recherche et le questionnaire utilisé dans la section 3. La section 4 présente l'analyse des résultats. Nous concluons dans la section 5.

## 2. Travaux Connexes

Les méthodes agiles sont parfois considérées comme des solutions miracles qui augmentent la productivité, diminuent les coûts et améliorent la qualité des projets (Chagas *et al.*, 2015) (Doraiaj *et al.*, 2010). Le Standish Group a révélé les défis de la mise en œuvre de projets indépendamment de l'utilisation des méthodes agiles ou non agiles (Hastie et Wojewoda, 2015). Pour lui les défis rencontrés dans la mise en œuvre de projets agiles sont plus faibles que ceux auxquels on doit faire face lors de l'utilisation de méthodes plus traditionnelles. Le groupe prévoit cependant la fin prochaine de la période faste des méthodes agiles pour une période où la gestion de projet sera beaucoup plus minimaliste (Johnson, 2020). Il faut noter que même si les équipes essaient d'appliquer correctement les principes de la méthodologie, elles constatent des problèmes pour la transition vers l'agile ou dans sa pratique. Beaucoup d'études existent sur les méthodes agiles et s'intéressent à certains facteurs clés, comme la communication, la confiance, le stress, etc.

**Communication.** La communication est un des facteurs humains les plus importants dans un projet agile et fait écho à l'un des principes du manifeste agile : *"The most efficient and effective method of conveying information to and within a development team is face-to-face conversation"* (Chagas *et al.*, 2015). L'amélioration de la communication est également le bénéfice le plus observé d'après l'enquête menée par (Begel et Nagappan, 2007) et (Pikkarainen *et al.*, 2008) suggèrent que les pratiques agiles ont des impacts positifs sur la communication entre les membres de l'équipe de développement. Ils montrent que les *daily meetings* permettent à chaque membre de

l'équipe de savoir sur quoi les autres membres sont en train de travailler et donc de mieux connaître l'état actuel d'avancement du projet.

**Confiance.** La confiance est un des facteurs les plus importants à la réussite d'un projet agile distribué (Dorairaj *et al.*, 2010) et l'usage des pratiques agiles peut renforcer la confiance entre les membres d'une équipe agile (Mchugh *et al.*, 2011). (Hasnain *et al.*, 2013) indique une différence importante dans la confiance entre les participants lors des jeux où la communication entre les participants était exigée. La confiance peut être construite en améliorant la communication (Dorairaj *et al.*, 2010), la responsabilisation des membres de l'équipe (Mchugh *et al.*, 2011) (Turk *et al.*, 2005), la transparence (Mchugh *et al.*, 2011), le partage de connaissances (Mchugh *et al.*, 2011) ou encore la compréhension des différences culturelles (Dorairaj *et al.*, 2010).

**Satisfaction.** Il y a une relation directe entre l'utilisation des pratiques agiles et la satisfaction au travail (Tripp *et al.*, 2016) (Melnik et Maurer 2006) (Kropp *et al.*, 2018). Les résultats de l'étude comparative de (Melnik et Maurer 2006) suggèrent qu'il y a 2 fois plus de membres satisfaits de leur travail dans les équipes agiles que dans les équipes classiques. Le modèle de Tripp et al. (Tripp *et al.*, 2016) induit que les différentes pratiques agiles impactent plus ou moins positivement 5 perceptions caractéristiques du travail : le travail en autonomie, le feedback, la variété des compétences, l'identité de la tâche (la tâche accomplie implique un résultat identifiable) et la signification de la tâche (elle a un impact sur l'entreprise ou la société en général). Ces 5 perceptions aboutissent à une perception de satisfaction au travail. (Melnik et Maurer 2006) et (Kropp *et al.*, 2018) constatent que le niveau de satisfaction des méthodes agiles est lié au niveau d'expérience dans les pratiques agiles. D'après (Kropp *et al.*, 2018), plus les répondants ont un niveau avancé dans les méthodes agiles plus ils sont satisfaits de ces méthodes. (Melnik et Maurer 2006) montrent également que la possibilité d'influencer sur les décisions, l'opportunité de travailler sur des projets intéressants et la relation avec les utilisateurs sont des facteurs qui influencent davantage la satisfaction des membres d'une équipe agile que la satisfaction de la charge de travail, l'opportunité d'avancement ou la participation au succès de l'entreprise.

**Stress.** Le stress est un des indicateurs permettant d'appréhender le ressenti des développeurs. Dans (Venkatesh *et al.*, 2020) l'épuisement au travail des développeurs est étudié au travers de l'ambiguïté des rôles et des conflits entre les rôles des développeurs. Les auteurs concluent que l'utilisation des pratiques agiles favorise une perception plus claire et non ambiguë par les développeurs de leurs rôles ce qui réduit l'épuisement au travail. (Laanti, 2013) montre un lien entre la responsabilisation de l'équipe et le niveau de stress. Les membres des équipes agiles plus responsabilisées ont tendance à ressentir moins de stress que les membres des équipes traditionnelles. Au travers de leur étude sur le stress, Meier et al. (Laanti, 2013) constatent que la majorité des répondants de l'enquête ont un niveau de stress neutre sur les projets agiles. Pour les développeurs, ce sont surtout des facteurs techniques sur la qualité du logiciel comme une architecture bien conçue ou un faible nombre de défauts qui réduisent le stress.

**Équipe.** L'idée d'une équipe agile auto-organisée vient d'un principe du manifeste agile : "*The best architecture, requirements, and designs emerge from self-organizing teams*". (Dorairaj *et al.*, 2010) et (Moe *et al.*, 2008) montrent que la transition d'une

équipe traditionnelle à une équipe agile auto-organisée est difficile et que cela vient du fait que certaines pratiques agiles nécessitent un transfert d'une partie des responsabilités traditionnelles du manager aux membres de l'équipe. (Moe *et al.*, 2008) démontre que la haute spécialisation des membres de l'équipe et la division du travail qui en découle favorisent l'autonomie individuelle mais constituent les plus gros obstacles pour aboutir à une équipe auto-organisée. Les membres de l'équipe ont tendance à se concentrer sur leurs tâches individuelles au détriment de la communication, ce qui réduit l'autonomie interne de l'équipe. Cette observation peut cependant être nuancée : d'après (Hoda *et al.*, 2010), la culture Néo-zélandaise qualifiée comme individualiste n'a pas eu d'impact négatif sur l'auto-organisation des équipes.

**Défis de la Transition Agile.** (Conboy *et al.*, 2011) et (Gandomani *et al.*, 2014) s'accordent sur le fait que les principaux challenges de la transition agile viennent de la gestion de l'aspect humain des équipes. (Conboy *et al.*, 2011) souligne plusieurs challenges humains lors de la transition agile: la peur du développeur que les autres membres de son équipe voient ses faiblesses, son manque de compétences techniques, métier ou sociales, le fait qu'il faut être polyvalent et bon dans tous les domaines, le manque de motivation ou d'intérêt pour les pratiques agiles, la nécessité de prendre des décisions, la difficulté à évaluer sa conformité avec les pratiques agiles et l'absence de recrutement ou de formation adaptés aux projets agiles. (Gandomani *et al.*, 2014) identifient 4 catégories: les obstacles au changement, l'accélération au changement, la perception des gens par rapport au changement et les facteurs d'incitation. Pour (Hekkala *et al.*, 2017) il s'agit essentiellement de défis culturels et managériaux.

**Avantages des Pratiques Agiles.** Certains travaux distinguent deux catégories de bénéfices : externes et internes (Solinsi et Petersen, 2016). Les bénéfices externes concernent les résultats de performance (objectifs atteints, coûts, temps, qualité) et de la relation client. Les bénéfices internes concernent l'équipe projet comme les connaissances, la satisfaction des membres de l'équipe, la communication, la coopération ou l'adaptabilité. (Vijayarathy *et al.*, 2008) montre que la flexibilité et la possibilité de délivrer des produits de meilleure qualité sont des bénéfices non négligeables. Les résultats de l'étude montrent aussi que la réduction des coûts et la réutilisation de code sont des bénéfices secondaires des pratiques agiles.

**Limites des Pratiques Agiles.** Pour (Turk *et al.*, 2002) il y a six limitations liées aux méthodes agiles: le manque de prise en charge des environnements de développement distribués, le soutien limité à la sous-traitance, la prise en charge limitée de la création d'artefacts réutilisables, le soutien limité au développement impliquant de grandes équipes, le soutien limité pour le développement de logiciels critiques pour la sécurité et le soutien limité au développement de gros logiciels complexes. Mais (Agrawal *et al.*, 2016) démontre que les limitations les plus importantes sont le manque de planification préalable et le manque de documentation suffisante. Une autre enquête montre d'autres soucis comme le fait que les méthodes agiles ne sont pas adaptées aux équipes projet de grande taille ou qu'il y ait trop de réunions (Begel et Nagappan, 2007).

L'intérêt de notre travail est d'avoir un recul plus important que certains travaux plus anciens pouvant dater d'avant les années 2010 et de se recentrer davantage sur la perception des développeurs que des managers ou des autres parties prenantes du projet.

### 3. Protocole de Recherche

Nous avons suivi la méthodologie proposée par Kitchenham (Kitchenham et Pfleeger, 2008) et composée des étapes suivantes : (a) Définition des questions de recherche (section 3.1), (b) Élaboration du questionnaire (section 3.2), (c) Obtention des données (section 3.3), (d) Analyse des données (section 4).

#### 3.1. Définition des Questions de Recherche

Nous avons défini notre question de recherche dans l'introduction : **QR : Comment sont perçues les méthodes agiles par les développeurs aujourd'hui?** Ce travail s'articulera autour de trois hypothèses principales identifiées en regard des travaux connexes de la littérature.

**H1 : Le développeur qui passe en mode agile doit adapter ses savoir-être et ses savoir-faire.** Cette première hypothèse permettra de s'intéresser à la façon dont les développeurs se sont adaptés à une méthode de travail agile et à comment ils ont vécu leur transition d'un mode de travail traditionnel à un mode de travail agile. Dans notre cas, les **savoir-faire** correspondront aux connaissances métier qui permettent de développer un projet informatique, les **savoir-être** correspondront aux comportements à adopter dans un environnement de projet agile.

**H2 : Le développeur voit les méthodes agiles comme un accélérateur à la réussite des projets.** Dans cette deuxième hypothèse, on étudie si les méthodes agiles favorisent davantage la réussite des projets que les méthodes traditionnelles et quels sont les facteurs favorisant la réussite des projets. On utilisera les critères classiques de coûts, de délais et de qualité pour mesurer cette réussite. L'intérêt est d'avoir la vision des développeurs sur l'impact des méthodes agiles sur la réussite des projets.

**H3 : Les méthodes agiles ont simplifié le travail du développeur au quotidien.** Dans cette troisième hypothèse, on s'interrogera sur comment les développeurs perçoivent leur position au sein d'un projet agile, si leur travail est simplifié, leur satisfaction au travail ainsi que sur les effets sur le stress des pratiques agiles.

#### 3.2. Élaboration du questionnaire

**Définition des Objectifs du Questionnaire.** L'enquête par questionnaire est une méthode de recherche plus quantitative que qualitative contrairement à l'enquête par entretien. L'intérêt ici est donc de permettre d'obtenir un plus grand nombre de réponses pour en déduire des tendances. Le questionnaire a été construit dans le but de valider ou invalider les hypothèses de la question de recherche.

**Choix de la Population Cible.** Ce questionnaire a été élaboré pour cibler les développeurs francophones. Il s'adresse aux développeurs travaillant dans des entreprises de toutes tailles, de la PME/startup aux grands groupes. Le questionnaire s'intéresse aussi bien aux personnes novices qu'aux personnes expérimentées en agilité. Notre principale restriction est de se concentrer sur les répondants ayant un rôle de développeur dans leur organisation, cependant nous avons conservé les réponses des autres types de répondants pour pouvoir effectuer une comparaison entre les réponses

des développeurs et les autres. Le questionnaire est resté ouvert un mois et diffusé sur plusieurs réseaux sociaux professionnels (*linked-in* et réseau interne) ainsi qu'à quelques personnes ciblées étant dans le profil.

**Choix du Format de Collection des Données.** Le questionnaire a été conçu dans le but d'obtenir des réponses à un peu moins de 40 questions sans décourager le répondant. La plupart des questions sont des questions fermées avec une seule réponse possible et aucune question n'est obligatoire pour ne pas bloquer un répondant. Les réponses au questionnaire sont anonymes, aucune coordonnée n'est requise pour participer. L'objectif est de faciliter la participation au questionnaire et de permettre aux répondants d'exprimer librement leur avis sur les méthodes agiles.

**Questionnaire.** Le questionnaire (cf. Tableau 1) est construit en 4 parties. On retrouve dans la première partie des questions pour récupérer des informations sur le profil des répondants. Ces informations seront utiles pour pouvoir dégager des tendances en les croisant avec les autres parties du questionnaire. Ensuite, chacune des trois autres parties correspond à l'une de nos trois hypothèses.

### 3.3 Collecte des Données.

Le questionnaire a été lancé sur plusieurs réseaux sociaux et nous avons obtenu 53 réponses valides dont plus de la moitié pour notre population cible de développeurs (développeurs et lead-développeurs).

## 4. Analyse des Résultats du Questionnaire

### 4.1 Questions Générales

**Âge.** La majorité des répondants sont relativement jeunes avec plus de 2 tiers entre 18 et 35 ans. Les 36-45 ans sont 15,4% et les plus de 45 moins de 10%.

**Type de Structure.** Plus de la moitié des répondants (56,6%) travaillent dans des grandes entreprises. On trouve 18,9% de répondants travaillant dans des entreprises de taille intermédiaire. Les répondants travaillant dans des petites structures comme les PME et start-ups sont moins nombreux avec moins de 10% des réponses.

**Poste.** Plus de la moitié des répondants occupent un poste de développeur ou lead développeur et constituent notre population cible. 15,4% sont *product owners* et une partie des répondants (28,8%) occupent un poste différent des réponses proposées.

**Durée de Temps de Travail sur les Méthodes Agiles.** La majorité des répondants sont peu expérimentés avec les méthodes agiles. Près de 2 tiers des répondants ont moins de 3 ans d'expérience et 34,6% des répondants ont moins d'un an d'expérience avec les méthodes agiles. 21,2% des répondants ont une expérience intermédiaire de 3 à 5 ans et ceux ayant plus de 5 ans d'expérience représentent moins de 15%.

**Formation sur les Méthodes Agiles.** La majorité des répondants ont reçu une formation sur les méthodes agiles mais près d'un tiers n'a pas été formé.

Tableau 1 : Questionnaire

<b>Profil</b>
Dans quelle tranche d'âge vous situez-vous ?
Dans quel type de structure travaillez-vous?
Quel est votre poste dans l'organisation ?
Depuis combien de temps travaillez-vous avec une méthode agile?
Avez-vous reçu une formation sur les méthodes agiles ?
Avez-vous déjà travaillé sur un projet avec une méthode traditionnelle non agile (cycle en V, en cascade, ...)?
Si oui, depuis combien de temps avez-vous travaillé avec ces méthodes traditionnelles?
<b>Hypothèse 1</b>
De combien de membres était composée votre dernière équipe projet agile (hors Scrum Master / PO) ?
Trouvez-vous que la transition vers les méthodes agiles a été simple pour vous ?
Pensez-vous que la méthode agile requiert une discipline plus rigide que les méthodes traditionnelles ?
Votre temps consacré aux réunions est-il plus important en quantité dans un projet agile ou dans un projet traditionnel ?
Comment vous êtes-vous adapté pour participer aux différentes réunions/cérémonies de projets agiles ?
Lorsque vous êtes sur un projet agile, devez-vous travailler avec moins de documentation, spécification que lors d'un projet traditionnel ?
Lorsque vous êtes sur un projet agile, avez-vous tendance à communiquer davantage avec les autres membres de votre équipe pour pallier à la potentielle limitation de documentation par rapport à un projet traditionnel?
Selon vous, est-ce que votre rôle dans un projet agile est plus spécialisé que dans un projet traditionnel ?
Lors d'un projet agile, à quelle fréquence êtes-vous amené à échanger avec votre manager ?
Lors d'un projet agile, pensez-vous être plus libre dans le choix des solutions pour le projet ?
<b>Hypothèse 2</b>
Quel est le taux de succès global (respect des coûts, délais et qualité à la fois) des projets agiles auxquels vous avez participé ?
Comment votre capacité à délivrer des fonctionnalités de votre organisation a évolué après adoption des approches agiles ?
Comment les méthodes agiles ont-elles affecté le coût de vos projets ?
Selon vous, quels facteurs peuvent expliquer que les projets agiles sont plus souvent une réussite que les projets traditionnels ? (question ouverte)
Comment les méthodes agiles ont-elles affecté la durée de vos projets ?
Lors de vos projets agiles, avez-vous constaté un allongement de la durée des sprints initialement prévue ?
Selon vous, quelles seraient les raisons de l'allongement général de la durée des sprints des projets agiles observé dans les projets agiles dans les entreprises ? (question ouverte)
Quelles sont selon vous les limites des méthodes agiles ?
<b>Hypothèse 3</b>
Pensez-vous que les méthodes agiles améliorent la communication entre les membres de votre équipe projet ?
Pensez-vous que les méthodes agiles améliorent votre confiance envers les membres de votre équipe projet ?
Trouvez-vous que les méthodes agiles améliorent la collaboration avec vos clients ?
Quel est votre niveau de satisfaction à propos de l'utilisation des méthodes agiles dans vos projets ?
Êtes-vous plus satisfait de votre travail personnel lorsque vous travaillez sur un projet agile ?
Est-ce que les méthodes agiles sont stressantes ?
Comment évaluez-vous votre niveau de charge de travail dans les projets agiles vs les projets traditionnels ?
Selon-vous, le nombre de réunions dans un projet agile est (trop important / Pas assez / Satisfaisant):
Selon-vous, la durée des réunions dans un projet agile est (Idéale / Trop courte / Trop longue):
Selon-vous, la fréquence des réunions en projet agile est-elle une contrainte pour l'organisation de votre journée de travail ?
Selon-vous, la quantité de documentation dans un projet agile est (Trop importante / Pas assez importante / Satisfaisante):
Selon-vous, la potentielle limitation de documentation dans un projet agile est-elle une contrainte pour votre compréhension des fonctionnalités du projet à développer ?

**Méthodes Traditionnelles.** Plus des trois quarts des répondants ont déjà travaillé sur un projet non agile (77.4%). Parmi ceux-ci, 24,4% ont une expérience de moins d'un an, 29,3% de 1 à 3 ans, 9,8% de 3 à 5 ans, 17,1% ont une expérience de 5 à 10 ans et 19,5% ont une expérience supérieure à 10 ans sur des méthodes traditionnelles.

#### **4.2 H1 : Le développeur qui passe en mode agile doit adapter ses savoir-être et ses savoir-faire**

**Taille de l'Équipe.** La majorité des répondants évoluent dans de petites équipes. Plus de la moitié a été dans une équipe agile de 4-6 membres et 28,8% des répondants sont dans des équipes de maximum 3 membres. Les répondants dans des équipes agiles de plus de 6 membres sont plus rares et représentent moins de 20% des réponses. La majorité des entreprises, indépendamment de leur taille, privilégient des équipes agiles plutôt petites de 3 à 6 personnes.

**Transition Simple vers les Méthodes Agiles.** Plus de 85% des répondants ont trouvé la transition plutôt simple. On n'observe pas de différence particulière entre les réponses des différentes tranches d'âges ou des postes des participants en dessous de 45 ans. Plus surprenant, le fait que les répondants aient déjà reçu ou non une formation sur les méthodes agiles ne semble pas influencer sur leur perception de la difficulté de la transition agile. Les réponses des développeurs qui ont eu une formation sont très similaires à celles de ceux qui n'ont pas été spécialement formés aux pratiques agiles.

**Discipline Rigide.** Les répondants sont partagés sur la question de la discipline des méthodes agiles. La moitié des répondants trouvent que les méthodes agiles ne requièrent pas plus de discipline tandis que l'autre moitié des répondants trouvent au contraire qu'elles requièrent davantage de discipline que les méthodes traditionnelles. Le ratio est le même pour les catégories de développeurs. Sur cette question, plus les répondants sont expérimentés, moins ils ont tendance à considérer que les méthodes agiles sont plus rigides que les méthodes traditionnelles. L'expérience des répondants impacte leur ressenti sur cette exigence de discipline. En effet, pour une expérience de l'agile de moins 3 ans, la moitié constate une discipline plus rigide, contrairement aux répondants plus expérimentés qui ne les jugent pas plus rigides que les autres. Ici le ressenti des catégories de développeurs démarque du reste des répondants puisque la proportion est identique dans les deux cas, quel que soit leur niveau d'expérience.

**Temps Important Consacré aux Réunions.** Presque la moitié des répondants trouvent qu'ils consacrent davantage de temps en réunion lors des projets agiles. Les réponses des *product owners* et coach agiles sont partagées (50% de oui et 50% de non) tandis que dans la catégorie des développeurs on observe plus de 57% de oui.

**Adaptation de Planning.** Une grande majorité des répondants (67.9%) s'adapte aux différentes réunions agiles en réservant des créneaux dans leur planning personnel. 20,8% préfèrent rester davantage disponibles dans la journée pour échanger avec leurs collègues si besoin est et une minorité adapte sa charge de travail en conséquence pour participer aux réunions agiles. On constate que les coachs agiles ont tendance à rester davantage disponibles dans la journée si l'équipe exprime le besoin de faire un point que les répondants occupants d'autres postes. On constate également qu'aucun développeur n'est amené à adapter sa charge de travail par rapport aux réunions.



**Documentation Identique.** Un peu plus de la moitié des répondants affirment travailler avec au moins autant de documentation dans un projet agile que dans un projet traditionnel. Dans les grandes entreprises, structure la plus représentée, une petite majorité seulement utilise moins de documentation et de spécifications.

**Communication Accrue pour Pallier au Manque de Documentation.** La majorité échange bien davantage avec les autres membres de l'équipe lors d'un projet agile. Étonnamment, un peu plus de 30% n'est pas dans ce cas. Une grande majorité des développeurs et *product owners* communiquent davantage sur les projets agiles que sur les projets traditionnels afin de compenser la limitation de la documentation.

**Rôle Polyvalent.** Plus de la moitié a un rôle plus polyvalent dans un projet agile que dans un projet traditionnel mais un quart déclare ne pas l'être. Plus surprenant, 17.6% ont un rôle plus spécialisé. On observe qu'un peu plus de la moitié des développeurs sont plus polyvalents lors des projets agiles et un tiers qui n'observe pas de différence. Presque 10% des développeurs, au contraire, sont plus spécialisés dans un projet agile.

**Fréquence Accrue des Échanges avec le Manager.** 43.1% des répondants échangent davantage avec leur manager dans un projet agile pour plus de la moitié de développeurs. Une minorité de répondants constate moins d'échanges.

**Liberté dans le Choix des Solutions.** On observe que plus de la moitié des participants se sent plus libre dans le choix des solutions dans un projet agile. 18.9% n'observent pas de différence et 22.6% pensent que cette liberté dépend surtout des caractéristiques du projet. Étonnamment, une minorité (5.7%) se sent accélératrice à la réussite des projets. Les développeurs se sentent en grande majorité plus libres dans le choix des solutions dans un projet agile.

**Discussion.** L'étude de la littérature scientifique met en évidence le fait qu'une amélioration de la communication et de la confiance est un facteur clé dans le succès des projets agiles (Begel et Nagappan, 2007) (Dorairaj *et al.*, 2008). Elle suggère également que les équipes agiles doivent être auto-organisées et que les membres de l'équipe doivent être davantage responsabilisés (Dorairaj *et al.*, 2010) (Moe *et al.*, 2008). De la même manière, les résultats du questionnaire tendent à valider l'hypothèse H1 sur l'adaptation des savoirs êtres et des savoir-faire des développeurs. Ils montrent que les équipes agiles sont en majorité plus petites que dans les projets traditionnels et que la transition agile est dans l'ensemble simple même s'il peut y avoir quelques difficultés. L'analyse du questionnaire montre que les développeurs passent plus de temps en réunions, communiquent davantage avec l'équipe, utilisent moins de documentation, qu'ils sont plus polyvalents et se sentent un peu plus libres dans le choix des solutions dans les projets agiles que traditionnels. Les résultats de cette analyse montrent également que la formation ou non sur les pratiques agiles ne semble pas avoir d'influence sur la difficulté perçue de la transition vers l'agile chez les développeurs et que plus ils sont expérimentés avec les pratiques agiles moins ils ont tendance à considérer ces pratiques comme rigides.

#### ***4.3 H2 : Le développeur voit les méthodes agiles comme un accélérateur de la réussite des projets***

**Amélioration du Taux de Succès Global.** On remarque qu'une proportion non négligeable de plus de 30% de répondants n'a pas connaissance des indicateurs de succès (temps, budget, qualité) de leurs projets agiles - c'est le cas pour 45% des développeurs, ce qui est plus élevé que dans les autres postes. Un peu plus de 20% des répondants constatent un taux de succès de leurs projets agiles supérieur à 80% et près d'un tiers indique un taux de succès compris entre 60 et 80%. Un peu plus de 10% des répondants constatent néanmoins un taux de succès de leurs projets agiles inférieur à 60%. Autre observation, les réponses venant de petites structures se concentrent sur un taux de succès entre 60 et 80% sur leurs projets agiles tandis que les moyennes et grandes entreprises ont des réponses sur des taux situés à plus de 80% mais aussi dans une moindre mesure sur des taux de réussite à moins de 60%.

**Amélioration de la Capacité à Délivrer des Fonctionnalités.** La grande majorité affirme que la capacité à délivrer des fonctionnalités a évolué positivement après adoption des approches agiles. Près d'un tiers a observé une grande augmentation de cette capacité. Seulement 15.4% ne constatent pas de changement dans leur capacité à délivrer après adoption des approches agiles. En revanche, aucun répondant n'a constaté de baisse de capacité à délivrer dans leur organisation après adoption des pratiques agiles. Plus de 70% des développeurs s'accordent sur le fait que leur capacité à délivrer des fonctionnalités a augmenté en utilisant les méthodes agiles. Seul un peu moins de 8% des développeurs n'observent pas de changement.

**Impact sur le Coût des Projets.** Plus de la moitié des répondants n'ont pas cette information. Sur le reste, 21.2% des répondants n'observent pas vraiment de variation de coût dans les projets agiles par rapport au budget initial. 13.4% des répondants ont observé une augmentation du coût de leurs projets agiles par rapport au budget initial. 7.7% des répondants ont au contraire constaté une diminution des coûts des projets agiles de plus de 20% par rapport au budget initial. La très grande majorité des développeurs ne peuvent donner une réponse à cette question. En excluant les réponses des répondants qui ne savent pas, la 2ème réponse la plus populaire est « pas de différence notable » à 21% puis une augmentation des coûts de 20% par rapport au budget initial environ comme vu précédemment dans la présentation des résultats.

**Facteurs de Réussite.** Les facteurs de réussite qui ressortent le plus dans les réponses sont la flexibilité qui permet de mieux s'adapter aux changements, une amélioration de la communication que ce soit avec l'équipe ou les clients, une meilleure organisation du planning, un découpage incrémental des tâches qui permet d'avoir une meilleure visibilité sur l'avancement du projet, des retours plus fréquents des clients et de délivrer plus régulièrement des fonctionnalités. Les développeurs mettent surtout en avant l'amélioration de la communication, les feedbacks plus réguliers, les itérations courtes et le découpage des tâches dans les bénéfices apportés sur les projets.

**Impact Positif sur la Durée des Projets.** Un peu plus de la moitié des répondants affirment que les méthodes agiles ont affecté positivement la durée de leurs projets avec 40.4% qui ont observé un gain de temps de plusieurs semaines. À peine plus de 5% ont déclaré des pertes de temps. On observe que les pratiques agiles permettent de gagner

plusieurs semaines dans un peu plus de 40% des réponses dans les différentes structures d'entreprises. Peu de réponses indiquent une perte de temps liée aux pratiques agiles et aucune concernant les développeurs.

**Peu d'Allongements de la Durée des Sprints.** 18,9% des répondants observent régulièrement un allongement de la durée de leurs sprints, mais cela reste rare pour plus de la moitié des répondants et n'arrive jamais pour 22,6% des participants. Le type de structure n'influe pas du tout sur ce ressenti des répondants, contrairement à ce que l'on pourrait penser. Nous avons demandé aux répondants ayant constaté un rallongement des sprints quelles seraient pour eux les raisons de cet état de fait et la réponse qui revient le plus souvent est la mauvaise ou la sous-estimation de la durée des tâches à effectuer. D'autres raisons sont également avancées comme le fait que des processus d'entreprises entravent le bon déroulement des sprints, la complexité de la demande client ou la complexité technique, des problèmes d'organisation, des facteurs humains et le retard accumulé sur les précédents sprints. Pour les développeurs, les raisons principales sont la mauvaise estimation de la complexité des tâches à effectuer, les imprévus au cours du projet, le manque de communication avec le client, le manque de certaines compétences techniques ou encore la non-disponibilité des membres de l'équipe du fait des congés ou du turn-over.

**Limites des Méthodes Agiles.** En première position (28,8%) arrive le manque de documentation, puis les contraintes budgétaires et le nombre de réunions trop important (25%), le manque de planification préalable (21,2%), l'exigence en termes de formation des pratiques agiles et le fait qu'elles ne soient pas adaptées aux grandes organisations (19,2% ) et enfin le manque de prédictibilité (17,3%). On remarque que 13,5% de répondants n'observent pas de limite particulière aux méthodes agiles et que seulement 9,6% considèrent que le fait de ne pas suivre les cycles de développement classiques constitue une limite aux pratiques agiles. Aucune tendance particulière ne se dégage en fonction des postes occupés ou du niveau d'expérience des répondants.

**Discussion.** La littérature scientifique avait déjà mis en valeur que seul 1 des 3 critères pour mesurer la réussite des projets (la qualité) était amélioré de manière significative par les pratiques agiles (Solinsi et Petersen, 2016). Bien que dans la lecture scientifique l'amélioration des coûts et des délais soit peu représentative des méthodes agiles, les équipes agiles tirent de nombreux bénéfices des méthodes agiles. Parmi les bénéfices significatifs les plus communs, on a l'amélioration de la communication interne de l'équipe (Begel et Nagappan, 2007), la flexibilité ou la capacité à apporter une réponse au changement (Vijayarathy et Turk, 2008), l'amélioration de la qualité du produit (Vijayarathy et Turk, 2008), la capacité de délivrer plus fréquemment (Vijayarathy et Turk, 2008), l'augmentation des feedbacks sur le produit et l'amélioration de la relation client. Cependant, pour tirer pleinement profit des bénéfices apportés par les pratiques agiles, il faut s'assurer que la transition vers agile soit bien menée en gérant les facteurs humains (Conboy *et al.*, 2011) (Gandomani *et al.*, 2014) et mettre en place l'ensemble et non une partie des pratiques d'une méthode agile. Les projets agiles en entreprises ont tendance à n'utiliser qu'une partie des pratiques agiles, mais il a été observé que plus une personne est expérimentée dans les méthodes agiles, plus elle aura tendance à utiliser davantage de pratiques de ces méthodes agiles (Kropp *et al.*, 2018). La littérature identifie également des limites aux méthodes agiles (Turk *et al.*, 2002)

(Agrawal *et al.*, 2016). L'analyse des résultats du questionnaire tend à invalider l'hypothèse 2 impliquant que les méthodes agiles seraient un accélérateur à la réussite des projets. Il est à noter que peu de développeurs connaissent les coûts, ou les taux de succès de leurs projets et ne peuvent donc réellement se prononcer sur ce critère. On observe seulement un taux de succès moyen des projets agiles plus important dans les moyennes/grandes entreprises que dans les petites entreprises. Un peu plus de la moitié des répondants constatent un gain de temps sur les projets agiles d'au moins quelques semaines. Cependant, la majorité des développeurs constatent une amélioration de leur capacité à délivrer avec les pratiques agiles. Les développeurs voient comme principaux facteurs de succès des projets agiles: l'amélioration de la communication, les feedbacks plus réguliers, les itérations plus courtes et le découpage des tâches. Les principales limites identifiées des pratiques agiles par les répondants sont le manque de documentation, les contraintes budgétaires, le nombre de réunions trop important et le manque de planification.

#### **4.4 H3 : Les méthodes agiles ont simplifié le travail du développeur au quotidien**

**Communication Améliorée.** Une amélioration est constatée pour plus des trois quarts des répondants, avec même une forte amélioration pour plus de la moitié. Seulement 7.5% n'observent pas réellement de différence avec ou sans les pratiques agiles sur la communication de l'équipe projet. Cette amélioration est constatée quelle que soit la taille de l'entreprise et quel que soit le poste occupé.

**Confiance Améliorée.** Les réponses sur la confiance envers les membres de l'équipe sont similaires à la précédente question sur la communication. Plus des trois quarts des répondants ont vu leur confiance envers les membres de leur équipe projet améliorée par les pratiques agiles avec même une grande amélioration pour la moitié des répondants (mêmes résultats pour la population des développeurs). Cette tendance est plus élevée dans les entreprises de petite ou moyenne taille, mais inversée pour les grandes entreprises et les start-ups où seulement respectivement 40% et 25% déclarent faire beaucoup plus confiance à leurs équipes. 20% des répondants de grandes entreprises ne distinguent pas de différence notable.

**Collaboration Client Améliorée.** Les réponses montrent que près de 85% des participants trouvent que les pratiques agiles améliorent la collaboration avec les clients. Plus de la moitié des répondants constatent une amélioration accrue de la collaboration client. Il est intéressant de noter que ce sont surtout les entreprises de taille intermédiaire qui semblent, à au moins 90%, connaître une grande amélioration de leur collaboration avec les clients. Les réponses indiquent que les pratiques agiles améliorent beaucoup la collaboration avec les clients quel que soit le rôle dans l'équipe. Les répondants de chaque poste sont au moins à 45% à penser que les méthodes agiles améliorent grandement la collaboration avec le client.

**Satisfaction Constatée.** Près de 85% des répondants sont satisfaits de l'utilisation des méthodes agiles dans leurs projets avec même un quart de très satisfaits, notamment les développeurs et les *products owners*. Plus de la moitié des répondants sont également plus satisfaits de leur travail personnel lors d'un projet agile, même si 28,8% des répondants n'observent pas réellement de changement.

**Stress.** Les avis sont très partagés. Près de 65% des répondants trouvent les méthodes agiles au moins un peu stressantes, avec 15,1% qui les trouvent même très stressantes. Au contraire des 35.8% qui ne les trouvent pas stressantes du tout. Il est intéressant ici de comparer ces ressentis pour les différents types de postes des répondants. Près de 50% des développeurs ne trouvent pas les pratiques agiles stressantes, avec seulement un peu moins de 9% des développeurs qui les trouvent très stressantes. Plus surprenant, concernant les répondants de la catégorie 'Autre', 26% trouvent les pratiques agiles très stressantes, ce qui est bien plus en proportion que les développeurs. On observe également que plus les répondants ont d'expérience avec les pratiques agiles, plus ils ont tendance à percevoir les méthodes agiles non stressantes, ce qui tendrait à dire que la pratique amène l'habitude et enlève le stress.

**Charge de Travail Identique.** Près de la moitié des répondants (47,1%) ne constatent pas de changement dans leur charge de travail. L'autre moitié constate une incrémentation de leur charge de travail (un peu pour 40% mais très importante pour 9.8%). Seuls 13,7% trouvent, qu'au contraire, leur charge de travail est moins importante dans les projets agiles. En ce qui concerne les développeurs, ils estiment majoritairement ne pas avoir plus de travail (73%).

**Nombre de Réunions Satisfaisant.** Près de 59% des répondants trouvent le nombre de réunions satisfaisant mais 39,2% trouvent qu'il y en a trop. On remarque qu'il s'agit essentiellement des développeurs qui trouvent qu'il y a trop de réunions en projet agile puisque la majorité des développeurs trouvent le nombre de réunions trop important.

**Durée des Réunions.** Plus de 67% des répondants trouvent la durée des réunions idéale. Elle serait trop longue pour 25% et trop courte pour 7.7%. Les développeurs rejoignent ici les avis des non développeurs. Seulement 23% des développeurs trouvent les réunions trop longues en projet agile.

**Contrainte de la Fréquence des Réunions.** Pour 93% cela a été une contrainte au moins une fois. Néanmoins, la moitié trouvent que cette fréquence est rarement une contrainte. Seuls 7,7% des répondants n'éprouvent jamais de difficulté avec les réunions agiles dans leur organisation personnelle. Les résultats montrent que les développeurs trouvent majoritairement que c'est rarement une contrainte, ce qui rejoint plus la tendance des autres postes.

**Documentation.** Plus de la moitié des répondants trouvent satisfaisante la quantité de documentation dans un projet agile mais 42,3% des répondants pensent qu'il n'y en a pas assez. Seuls 5,8% des répondants trouvent qu'il y en a trop. Les avis sont assez partagés sur l'ensemble des postes des répondants. Il y a autant de personnes qui pensent que la quantité de documentation est satisfaisante que de personnes qui pensent qu'elle est insuffisante chez les développeurs. Les répondants d'entreprises de taille intermédiaire sont plus de 75% à trouver la quantité de documentation insuffisante. Les répondants en grandes entreprises sont plus partagés avec 43% qui trouvent la quantité insuffisante tandis que 46% la trouvent suffisante. Dans les autres structures d'entreprises, au moins 75% des répondants sont satisfaits de la quantité de documentation dans leurs projets agiles. Plus de 85% des répondants ont éprouvé des difficultés dans la compréhension des fonctionnalités du projet à développer dû à la limitation de la documentation et c'est une contrainte pour la majorité des participants.

Seuls 13,5% des répondants n'ont jamais perçu la limitation de documentation comme une contrainte. Les chiffres sont sensiblement les mêmes quel que soit le poste occupé.

**Discussion.** Les études montrent qu'il y a une relation positive entre l'usage des pratiques agiles et la satisfaction au travail. Les membres d'une équipe utilisant les pratiques agiles sont plus satisfaits que ceux d'une équipe traditionnelle (Melnik et Maurer, 2006). Il a également été observé un lien entre le niveau de satisfaction et le niveau d'expérience (Kropp *et al.*, 2018). Plus les membres d'une équipe ont un niveau avancé dans les méthodes agiles et plus ils en sont satisfaits. La littérature indique également que les membres responsabilisés d'une équipe agile sont moins stressés (Meier *et al.*, 2018). Une étude montre que l'usage des pratiques agiles favorise une perception plus claire et non ambiguë par les développeurs de leurs rôles ce qui réduit l'épuisement au travail (Venkatesh *et al.*, 2020). L'hypothèse H3 sur la simplification du travail du développeur au quotidien est validée dans l'ensemble par les résultats du questionnaire. Les résultats montrent que 100% des développeurs sont satisfaits de l'utilisation des méthodes agiles. 65% des répondants trouvent les méthodes agiles un peu stressantes, mais l'analyse montre que chez les répondants expérimentés de 3 à 5 ans sur les pratiques agiles, 75% d'entre eux ne trouvent pas les méthodes agiles stressantes. La majorité des développeurs ne voient pas réellement de changement dans leur charge de travail en projet agile par rapport en projet traditionnel. Cependant, un peu plus de la moitié des développeurs trouvent le nombre de réunions trop important et la grande majorité trouvent que la limitation de documentation induite constitue une contrainte pour la compréhension des fonctionnalités à développer.

#### 4.5 Obstacles à la Validité du Questionnaire

La recherche qualitative est basée sur des données subjectives, interprétées et contextuelles. (Thomson, 2011) propose 5 catégories de validité.

La **validité descriptive** se réfère à l'exactitude des données. Nous avons unifié les critères utilisés dans l'étude et structuré l'information à recueillir à l'aide d'un formulaire d'extraction de données pour un enregistrement uniforme des données.

La **validité théorique** dépend de la capacité d'obtenir l'information que l'étude est censée saisir. Nous avons créé un questionnaire spécifiquement adapté aux développeurs en agilité, chaque question étant directement reliée à l'une de nos trois hypothèses de recherche. Le questionnaire a été publié sur les réseaux sociaux et nous n'avions pas de possibilité de vérifier si les répondants correspondaient totalement au profil demandé, ce point étant donc laissé à leur libre appréciation. Nous avons tout de même effectué un nettoyage des données pour supprimer les répondants ne correspondant pas du tout aux utilisateurs de méthodes agiles sur la base de leurs réponses.

La **validité de généralisation** concerne la capacité de généraliser les résultats. Le nombre de répondants au questionnaire n'est que de 53 et parmi les répondants près de la moitié ne sont pas développeurs. Il aurait été plus intéressant de pouvoir analyser davantage de réponses avec une proportion plus grande de développeurs qui sont la population cible. Autre faiblesse du questionnaire, les répondants sont en majorité assez jeunes, 2/3 des répondants sont âgés entre 18 et 35 ans. Les répondants de plus de 45

ans représentent un peu moins de 10% du total des répondants. Il aurait été intéressant d'avoir davantage de réponses de personnes plus âgées et expérimentées pour faire des comparaisons en fonction des catégories d'âges et d'ancienneté des répondants.

La **validité évaluative** est obtenue lorsque les conclusions sont raisonnables compte tenu des données. Deux chercheurs ont étudié les résultats et validé chaque conclusion.

La **validité de la transparence** fait référence à la répétabilité du protocole de recherche. Le protocole utilisé pour le questionnaire est suffisamment détaillé et le questionnaire est donné dans le tableau 1 pour permettre une répétition.

## 5. Conclusion

Nous avons présenté les résultats d'un questionnaire s'intéressant à la manière dont sont perçues les méthodes agiles par les développeurs aujourd'hui. Trois hypothèses ont été établies : H1 : Le développeur qui passe en mode agile doit adapter ses savoir-être et ses savoir-faire, H2 : Le développeur voit les méthodes agiles comme un accélérateur à la réussite des projets, H3 : Les méthodes agiles ont simplifié le travail du développeur au quotidien. Nous avons étudié la littérature scientifique sur les trois hypothèses et comparé les résultats déjà existants avec les conclusions de notre questionnaire. Ce travail nous a permis de valider les deux hypothèses H1 et H3 et d'invalider l'hypothèse H2 (avec la nuance que les développeurs n'ont que peu d'informations sur certains points). Cependant il serait intéressant d'obtenir plus de réponses à ce questionnaire pour avoir plus de facilité à généraliser les conclusions tirées. Il serait également utile de faire une étude plus qualitative sur le sujet en interviewant plusieurs experts, ce qui permettrait de valider ou invalider plus avant les hypothèses de ce travail.

## Bibliographie

- Agrawal, A., Atiq, M. A., Maurya, L. S. (2016). A current study on the limitations of agile methods in industry using secure Google Forms. *Procedia Computer Science*, vol. 78, 291-297.
- Anderson, D., *Agile Management for Software Engineering : Applying the Theory of Constraints for Business Results*, Prentice Hall 2003, 2003
- Augustine, S., Payne, B., Sencindiver, F., Woodcock, S. (2005) Agile project management: steering from the edges *Communications of the ACM*, Volume 48, Number 12, pp 85-89
- Beck, K. "Extreme Programming Explained", Addison-Wesley, 2000
- Beck, K., et al. (2001) *The Agile Manifesto*. Agile Alliance. <http://agilemanifesto.org/>
- Begel, A. & Nagappan, N. (2007). Usage and perceptions of agile software development in an industrial context: An exploratory study. In : *First International Symposium on Empirical Software Engineering and Measurement (ESEM 2007)*. IEEE. p. 255-264.
- Chagas, A., Santos, M., Santana, C., Vasconcelos, A. (2015). The impact of human factors on agile projects. In : *2015 Agile Conference*. IEEE. p. 87-91.
- Conboy, K., Coyle, S., Wang, X., Pikkarainen, M. (2011). People over process: key people challenges in agile development.
- Dorairaj, S., Noble, J., Malik, P. (2010). Understanding the importance of trust in distributed Agile projects: A practical perspective. In : *International Conference on Agile Software Development*. Springer, Berlin, Heidelberg. p. 172-177.

- Gandomani, T. J., Zulzalil, H., Ghani, A. A., Sultan, A. B., Sharif, K. Y. (2014). How human aspects impress Agile software development transition and adoption. *International Journal of Software Engineering and its Applications*, vol. 8, no 1, p. 129-148.
- Hasnain, E., Hall, T., Shepperd, M. (2013). Using experimental games to understand communication and trust in agile software teams. In : 2013 6th International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE). IEEE. p. 117-120.
- Hastie, Wojewoda. (2015) Standish Group 2015 Chaos Report - Q&A with Jennifer Lynch.
- Hekkala, R., Stein, M., Rossi, M., & Smolander, K. (2017). Challenges in transitioning to an agile way of working.
- Hoda, R., Noble, J., Marshall, S. (2010). Organizing self-organizing teams. In : 2010 ACM/IEEE 32nd international conference on software engineering. IEEE. p. 285-294.
- Johnson J. (2020) Standish Group – CHAOS 2020: Beyond Infinity
- Kitchenham B.A., Pfleeger S.L. (2008) Personal Opinion Surveys. In: Shull F., Singer J., Sjøberg D.I.K. (eds) *Guide to Advanced Empirical Software Engineering*. Springer, London.
- Kropp, M., Meier, A., Anslow, C., Biddle, R. (2018). Satisfaction, practices, and influences in agile software development. In : *Proceedings of the 22nd International Conference on Evaluation and Assessment in Software Engineering 2018*. p. 112-121.
- Laanti, M. (2013). Agile and Wellbeing--Stress, Empowerment, and Performance in Scrum and Kanban Teams. In : 2013 46th Hawaii International Conference on System Sciences. IEEE. p. 4761-4770.
- Mchugh, O., Conboy, K., Lang, M. (2011). Agile practices: The impact on trust in software project teams. *IEEE software*, vol. 29, no 3, p. 71-76.
- Meier, A., Kropp, M., Anslow, C., Biddle, R. (2018). Stress in agile software development: practices and outcomes. In : *International Conference on Agile Software Development*. Springer, Cham. p. 259-266.
- Melnik, G., Maurer, F. (2006). Comparative analysis of job satisfaction in agile and non-agile software development teams. In : *International conference on extreme programming and agile processes in software engineering*. Springer, Berlin, Heidelberg. p. 32-42.
- Moe, N. B., Dingsoyr, T., & Dyba, T. (2008). Understanding self-organizing teams in agile software development. *19th Australian conference on software engineering*. IEEE. p. 76-85.
- Pikkariainen, M., Haikarrar, J., Salo, O., Abrahamsson, P., Still, J. (2008). The impact of agile practices on communication in software development. *Empirical Software Engineering*, vol. 13, no 3, p. 303-337.
- Schwaber K., *SCRUM Development Process*, 10th Annual ACM Conference on Object Oriented Programming Systems, Languages, and Applications (OOPSLA), 1995.
- Solinsi, A. & Petersen, K. (2016). Prioritizing agile benefits and limitations in relation to practice usage. *Software quality journal*, vol. 24, no 2, p. 447-482.
- Thomson, S. B.. (2011) *Qualitative Research: Validity*. JOAAG, Vol. 6. No 1
- Tripp, J. F., Riemenschneider, C., Thatcher, J. B. (2016). Job satisfaction in agile development teams: Agile development as work redesign. *Journal of the Association for Information Systems*, vol. 17, no 4, p. 1.
- Turk, D., France, R., & Rumpe, B. (2002). Limitations of agile software processes. In : *Proceedings of the Third International Conference on eXtreme Programming and Agile Processes in Software Engineering*. p. 43-46.
- Turk, D., Robert, F., Rumpe, B. (2005). Assumptions underlying agile software-development processes. *Journal of Database Management (JDM)*, vol. 16, no 4, p. 62-87.
- Venkatesh, V., Thong, J. YL, Chan, F. K., Hoehle, H., Spohrer, K. (2020). How agile software development methods reduce work exhaustion: Insights on role perceptions and organizational skills. *Information Systems Journal*, vol. 30, no 4, p. 733-761.
- Vijayasathy, L. E. O. R., Turk, D. (2008). Agile software development: A survey of early adopters. *Journal of Information Technology Management*, vol. 19, no 2, p. 1-8.



---

## Extraire et organiser des connaissances sur les pathogènes

### Projet EPICURE

**Leïla RENARD<sup>1</sup>, Thérèse LIBOUREL<sup>2</sup>, Catherine MOULIA<sup>2</sup>,  
Laurent GAVOTTE<sup>2</sup>**

1. Université de Montpellier, France

*leilarenard.freelance@gmail.com*

2. ESPACE-DEV, Univ Montpellier, IRD, Univ Antilles, Univ Guyane, Univ Réunion,  
Montpellier, France

*therese.libourel@umontpellier.fr, catherine.moulia@umontpellier.fr, laurent  
.gavotte@umontpellier.fr*

---

**RÉSUMÉ.** *L'article présente l'approche suivie au sein du projet Epicure. Le projet consiste en la mise en œuvre d'un système d'information dédié, véritable base de connaissances approfondie et élargie nécessaire à l'étude et/ou la gestion du risque infectieux et parasitaire de notre monde. Le plan général du projet sera présenté avant de détailler le cœur de notre propos i.e la phase de conceptualisation qui se fonde sur l'interdisciplinarité et a pour objectif de remettre la connaissance au cœur de la science.*

**ABSTRACT.** *The article presents the approach taken within the Epicure project. The project consists in the implementation of a dedicated information system, genuine extensive and expanded knowledge base necessary for the study and/or management of the infectious and parasitic risk of our world. The general plan of the project will be presented before detailing the heart of our proposal i.e the conceptualisation phase, which is based on interdisciplinarity and aims to put knowledge back in the core of science.*

**MOTS-CLÉS :** *Pluridisciplinarité, Interdisciplinarité, Connaissance, Pathogène, Épidémiologie*

**KEYWORDS:** *Pluridisciplinarity, Interdisciplinarity, Knowledge, Pathogen, Epidemiology*

---

## 1. Introduction

Notre travail s'inscrit dans le cadre du projet EpiCURE (*Eco-Epidemiology of Animal and Human Pathogens Comprehensive and Utilitary Resources*) développé grâce à un financement de l'I-site MUSE (Montpellier Université d'Excellence).

Ce projet, à visée pédagogique, s'appuie sur la nouvelle mention de master de l'université de Montpellier « Eco - EPI : Eco-épidémiologie », formation centrée spécifiquement autour des agents pathogènes des humains et des animaux.

Il s'agit de mettre en place une plateforme permettant de recueillir et rassembler des « informations exhaustives » concernant les agents pathogènes de maladies d'intérêt en santé humaine et vétérinaire, afin de les synthétiser, les présenter et permettre leur analyse de manière globale. L'ensemble des informations disponibles sera issu de la littérature scientifique ainsi que de sources gouvernementales ou supra-gouvernementales validées (OMS, OIE etc.) pour chacun des pathogènes connus. Les références des articles seront conservées par l'intermédiaire de métadonnées dont le noyau sera conforme aux standards classiques.

Par ensemble d'informations disponibles, nous entendons les données médicales et vétérinaires, biologiques, d'épidémiologie descriptive, mais aussi éco-environnementales et socio-économiques. Si une partie de ces informations sont actuellement disponibles via d'autres bases de données<sup>1</sup>, elles se présentent sous forme fragmentée et incomplète. De plus, certaines de ces plateformes ne sont accessibles ni au grand public, ni à tous les acteurs scientifiques.

Ce projet présente 4 originalités :

- Il se positionne clairement dans l'approche « One Health » (Zinsstag *et al.*, 2020), (Evans *et al.*, 2014) ou « Une seule santé », vision holistique de la santé et des liens entre santé, qualité de l'environnement (eau, air...), climat, socio - et éco-systèmes.
- Cette pluridisciplinarité structurelle du « One Health » est complétée par la mise en œuvre concrète de la pluridisciplinarité et de l'interdisciplinarité via les échanges nécessaires entre informaticiens et éco-épidémiologistes (Bizouarn, 2016).
- Le projet se construit par des échanges recherche-enseignement qui demandent de penser la pédagogie autour de la participation active des étudiants et de la co-construction. Brièvement, les informations seront collectées par les étudiants des promotions successives de la nouvelle mention, puis validées par l'équipe pédagogique constituée d'experts des différents champs thématiques et disciplinaires concernés, et finalement intégrées dans la plateforme avant d'être mises à disposition de la communauté scientifique, universitaire ou du grand public.

1. Quelques exemples (Blanchet *et al.*, 2019) Plateforme IFB regroupant diverses collections. (Plateforme, SPF, s. d.) Plateforme ESA (Goehrs *et al.*, 2012) Portail Epidémiologie - France | Health Databases

– Enfin, il s’inscrit dans un contexte de « sciences ouvertes » (Kembellec, 2019), les connaissances explicites devant enrichir tout à la fois un cursus universitaire et répondre aux questions des spécialistes comme à celle du grand public intéressé (Clément, 2004).

La suite de l’article est organisée de la manière suivante : dans un premier temps, nous présenterons dans la section 2. *Planification et méthodologie globale du projet* la méthodologie globale du projet, les premières interactions entre éco-épidémiologistes et informaticiens ayant permis de mettre en œuvre un plan de capitalisation des connaissances. Les maladies parasitaires et infectieuses sont ainsi abordées de manière « systémique » en nous focalisant sur les aspects structurels (organisation) et les aspects fonctionnels (fonctionnement dont découle le concept de processus). Dans la section 3. *Les résultats actuels*, nous détaillerons comment diverses avancées ont été réalisées et illustrerons les résultats obtenus au travers d’exemples. La section 4. *Spécification de la plateforme* décrira brièvement les composants utilisables dans la partie opérationnelle envisagée. Enfin la section 5. *Conclusion* présentera les propositions ultérieures.

## 2. Planification et méthodologie globale du projet

Le projet est structuré en trois grandes phases : la conceptualisation des divers « domaines d’intérêt » sous leurs différents aspects, la spécification de la plateforme et la réalisation opérationnelle.

Nous ne nous attarderons dans l’article que sur la première phase.

Retenons tout d’abord, les quatre axes thématiques qu’Epicure se propose de traiter car ils permettent d’aborder l’éco-épidémiologie dans toute sa complexité :

– En premier lieu la plateforme comprendra un corpus de données biologiques à différentes échelles et niveau d’analyse. Celui-ci inclut la taxonomie des agents pathogènes et les éléments de biologie des organismes en fonction de cette taxonomie. Il s’agira donc de données concernant la morphologie, la reproduction, la structure du génome et de la cellule, ou encore les mécanismes infectieux. On ajoutera également des données sur les connaissances du génome des agents pathogènes (liens vers les banques de gènes, gènes d’intérêts, éléments spécifiques de génome, diversité génétique), de son évolution ainsi que sur la biodiversité intra-taxon et la génétique des populations des agents pathogènes. Ce corpus inclut également les informations d’écologie parasitaire, c’est-à-dire le cycle parasitaire naturel (sylvatique, domestique, synanthropique) indiquant les espèces hôtes intermédiaires et définitives, les organes infectés et les modalités d’exploitation associés du milieu hôte. Le corpus inclut enfin les informations de bio-géographie, c’est-à-dire la répartition des agents pathogènes à l’échelle mondiale ainsi que les zones et les dates d’introduction récentes.

– La plateforme comprendra également les données épidémiologiques *stricto sensu*, informations descriptives relatives aux différentes maladies telles que le nombre de reproduction de base (R0), la prévalence, l’incidence, la létalité, le caractère épidé-

mique et/ou endémique, la stratification dans la population hôte en fonction de divers paramètres (âge, statut socio-économique). Ces données seront renseignées en l'absence et en présence de mesures de gestion ou de contrôle lorsque ces dernières sont disponibles.

– L'axe 3 comprendra les données sur le contexte socio-économique impliqué dans l'épidémiologie, incluant entre autres les pratiques et usages culturels, les filières d'élevage, les échanges marchands, l'organisation du système de santé.

– Enfin, des données de santé constitueront l'axe 4. Elles incluent le tableau clinique, les éléments et outils de diagnostic (avec des indications de performance et d'accessibilité des tests), la prophylaxie, et les traitements existants. Les axes 3 et 4 sont joints par la notion de gestion (incluant les mesures de prévention, de surveillance et de contrôle non médicamenteux).

La conceptualisation est donc le processus de représentation de la connaissance extraite. Elle est réalisée par une construction collaborative, ce qui est un vrai défi compte-tenu de la complexité. Elle est abordée par modélisation conceptuelle basée sur l'approche objet.

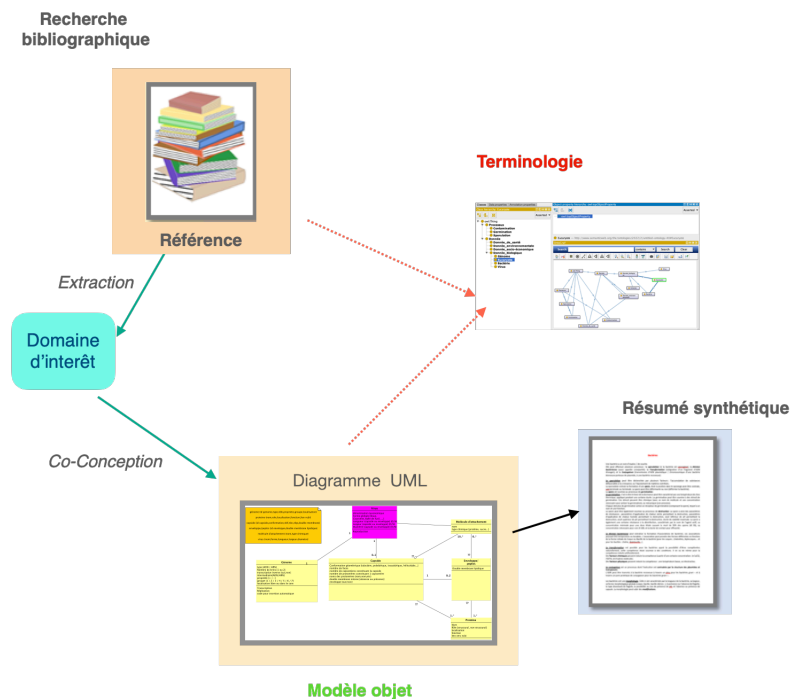


FIGURE 1. Vue globale de l'étape de conceptualisation

Le point de départ est donc la recherche bibliographique qui s'effectue dans un ensemble de sources textuelles issues de revues, manuels, articles scientifiques, sites Web. Seront extraits, de la source référence sélectionnée, par analyse textuelle, un en-

semble de termes relatifs à l'intérêt du lecteur c'est-à-dire ceux relatifs à une partie de la terminologie concernant la donnée biologique ou épidémiologique (nous reviendrons sur l'intérêt de cette terminologie dans la section 5).

La source référence sélectionnée sera conservée en métadonnée des données ciblées.

La modélisation objet est effectuée par consensus interdisciplinaire sous forme de diagrammes de classes exprimés en formalisme UML *Unified Language Modeling* (Muller, Gaertner, 2000), langage normé et universel (voir la Figure 2).

Le processus général de construction des diagrammes UML a nécessité de nombreux aller-retours entre différentes étapes de travail commun pour atteindre un consensus ;

- L'analyse de texte de chaque référence se fait par classification des termes (substantif, verbe, qualificatif) puis par abstraction sous forme d'éléments (classes, attributs, relations) de diagrammes UML de légende commune (voir la Figure 3) rendant les connaissances sous-jacentes plus simples (mais justes) et lisibles.

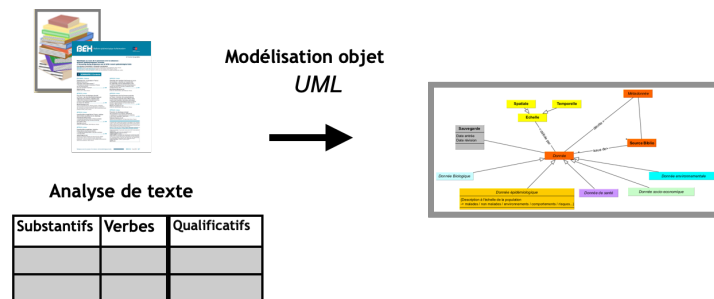


FIGURE 2. *Conceptualisation Modélisation objet*

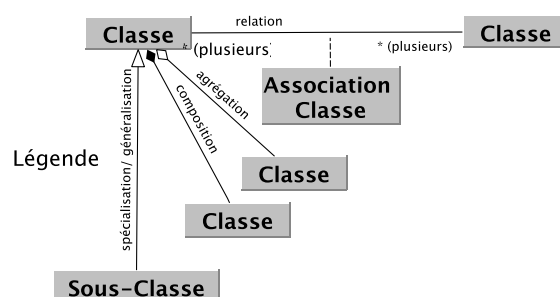


FIGURE 3. *Légende formalisme UML*

- Les divers aller-retours effectués sur ces analyses nous ont permis soit de regrouper les informations répertoriées à des niveaux d'abstraction plus généraux, soit à compléter et préciser certains éléments.

– Chaque diagramme UML obtenu est ensuite décrit par un texte résumé. Celui-ci est écrit de sorte à ce que tous les concepts soient explicités et que les liens entre eux soient mis en évidence. Ils permettront aux collaborateurs actuels et futurs du projet de facilement discuter et réviser le contenu des diagrammes, s’assurer qu’aucune erreur ou qu’aucun oubli n’a été fait, ceci sans être familier avec le langage UML.

– A partir de ces résumés et des diagrammes déjà obtenus nous simplifierons les diagrammes afin de faciliter leur traduction sous forme de schémas de base de données.

### 3. Les résultats actuels

#### 3.1. Vue d’ensemble des diagrammes

Les informations répertoriées dans divers diagrammes relatifs aux premières analyses nous ont permis de concevoir une vue générale articulée autour des quatre domaines particuliers de données évoqués précédemment (cf. Figure 4).

Ce diagramme constitue alors un point de départ pour ensuite décliner l’ensemble des diagrammes UML plus détaillés (Par exemple, la classe « Virus » dans le diagramme général sera déclinée dans un diagramme plus spécifique). Il permettra de lier entre eux les divers diagrammes plus spécifiques une fois terminés, les liens d’association simple, agrégation, composition, spécialisation / généralisation entre les différentes classes à décliner étant déjà présents à cette étape.

A l’heure actuelle, nous avons structuré des concepts se rapportant aux données biologiques, incluant la taxonomie, et en fonction de celle-ci des données morphologiques et génomiques. Seront également décrits en détail les processus infectieux aux différentes échelles (de la molécule à l’organisme) ainsi que le cycle parasitaire dans son ensemble incluant leur spectre d’hôtes ainsi qu’un ensemble de données biologiques sur ces hôtes.

Voici une liste des différents diagrammes UML produits jusqu’à maintenant :

- « Méta-données »: associe toute donnée à un ensemble de métadonnées référant également les sources de référence.
- Schémas permettant la descriptions des différents organismes pathogènes : « virus », « bactéries », « plathelminthes », « eucaryotes unicellulaires ».
- « Cycle parasitaire »: permet de décrire l’ensemble des cycles parasites pouvant exister, quels que soient les hôtes et les agents pathogènes.
- « Taxonomie »: classe les agents pathogènes de façon hiérarchiques en fonction des taxons auxquels ils appartiennent.
- « Génome »: décrit l’organisation structurelle des génomes des eucaryotes, des bactéries et des virus.
- « Description du système des eucaryotes »: description de l’organisation hiérar-

chique d'un être vivant eucaryote.

- « Mode de transmission » (décliné en 5 schémas en fonction des hôtes contaminants et contaminés) : décrit les paramètres d'entrée des agents pathogènes dans un hôte.

- « Sortie des organismes »: décrit les paramètres de sortie des agents pathogènes d'un hôte.

- « Processus »: Formalisation de la notion de processus, impliquant 2 agents et étant soumis à plusieurs paramètres.

- « Contamination »: schéma simplifié représentant les entrées et sorties associées au processus de contamination.

- « Cycle parasitaire (dont contamination) »: schéma simplifié représentant le cycle parasitaire et incluant le schéma « Contamination ».

### 3.2. *Détail du processus de conceptualisation de certains diagrammes*

Nous allons détailler dans un ordre « logique » le processus de conceptualisation de diagrammes UML ayant particulièrement impliqué l'interdisciplinarité. Grâce aux divers points de vue sur les exemples étudiés, nous avons progressivement rassemblé et unifié la vision en gagnant en généralité par abstraction successives. Cela a été réalisé pour les diagrammes explicitant les structures et ceux explicitant les processus (réifiés).

#### 3.2.1. *Diagramme de vue générale*

Le diagramme général (cf. Figure 4) a été initié par réflexion après le début de la construction des premiers diagrammes de description des organismes pathogènes. Nous avons eu besoin de le construire afin de le concevoir comme un **référentiel** permettant de visualiser l'ensemble des connaissances à sauvegarder dans la base, c'est-à-dire l'ensemble des éléments devant y figurer (structures et processus) ainsi que l'ensemble des liens les reliant.

Ce diagramme a été construit à partir des domaines des données du projet (données biologiques, épidémiologiques, de santé, socio-économiques, environnementales), en identifiant puis en organisant les différentes informations que devaient regrouper ces domaines.

Ce diagramme est modifié au fil du temps pour intégrer les nouveaux diagrammes qui n'avaient pas été envisagés lors de sa conception initiale. En effet, le projet étant très vaste, il est difficile de cerner l'intégralité des concepts à aborder.

#### 3.2.2. *Cycle parasitaire général*

A chaque nouvelle lecture, les extractions de connaissance et les réflexions nous amènent à créer de nouveaux diagrammes UML afin d'abstraire des caractéristiques déjà présentes dans divers exemples permettant ainsi une vision générique de celles-ci ; c'est la cas du cycle parasitaire général.

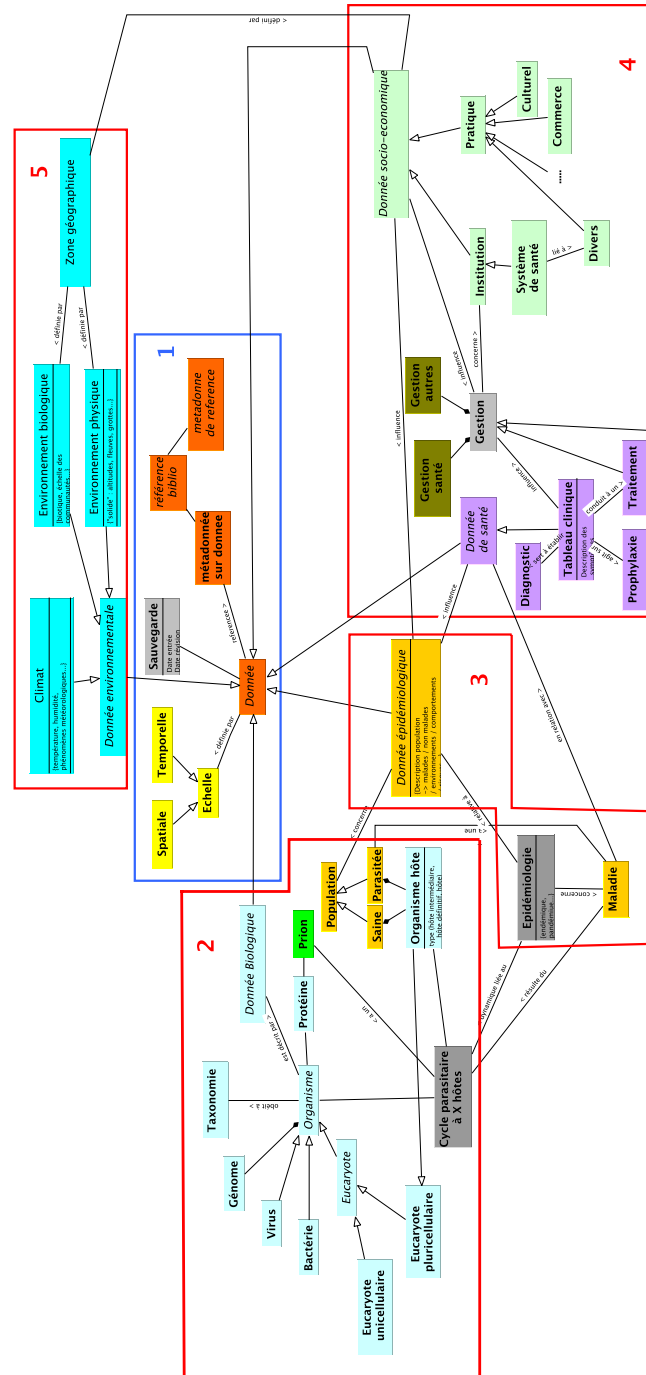


FIGURE 4. *Vue globale du diagramme conceptuel général, correspondant à la vue d'ensemble des données à sauvegarder dans la base de données du projet EpiCURE 1. 1.Données abstraites 2.Données biologiques 3.Données épidémiologiques 4.données de santé et socio-economique 5.données environnementales. Les classes grisées représentent des processus, les autres couleurs permettent de regrouper visuellement les éléments en fonction des thèmes abordés*



Les parasites sont des organismes vivant aux dépens des ressources d'un hôte. Leur cycle de vie, ou cycle parasitaire, comporte plusieurs stades évolutifs (par exemple: œuf, stades juvéniles, stade adulte) et peut se réaliser à l'intérieur d'un ou plusieurs compartiments hôtes, un compartiment hôte étant l'ensemble des organismes vivants dans lesquels un parasite peut effectuer une partie spécifique de son cycle. Les hôtes peuvent être intermédiaires (ils accueillent alors les stades larvaires effectuant éventuellement une multiplication asexuée), définitifs (ils accueillent alors le stade adulte effectuant la reproduction sexuée), ou facultatifs.

Le diagramme du Cycle parasitaire général (cf. Figure 5) a été construit après avoir conceptualisé les cycles parasitaires des plathelminthes et des eucaryotes unicellulaires. Les plathelminthes (des vers plats) ont une grande diversité de cycles parasitaires du fait notamment qu'en fonction des espèces, ils ont un nombre variable de stades évolutifs ainsi qu'un nombre variable d'hôtes nécessaires à leur cycle. De plus, ils se déplacent, se développent et se reproduisent dans leurs hôtes (Combes *et al.*, 2018) (Pereira *et al.*, 2013). Les eucaryotes unicellulaires (agent pathogène composé d'une seule cellule nucléée) ont également plusieurs stades évolutifs, un nombre d'hôtes variable, se déplacent, se développent et se reproduisent dans leur hôte (Bates, 2007) (Loftus *et al.*, 2005).

Nous avons alors remarqué les similarités entre les deux diagrammes de cycle parasitaire effectués pour ces deux types d'organismes, et avons en conséquence décidé de généraliser le cycle parasitaire en concevant un diagramme UML qui pourrait permettre de décrire les cycles parasitaires de l'ensemble des agents pathogènes. Pour ce faire, nous nous sommes également appuyés sur le cycle de l'Anguilliose (*Strongyloidiasis biology*, 2019) et de Anisakis (*Cycle évolutif de l'anisakiose*, 2016), deux nématodes (cf. Figure 6) ayant des cycles parasitaires très complexes, afin de minimiser la probabilité d'oublier des éléments nous permettant de les décrire.

Une fois le diagramme réalisé, nous avons testé le modèle avec différents agents pathogènes.

### 3.2.3. Mode de transmission

*Idée de départ* : Un agent pathogène peut être transmis d'un hôte à un autre de multiples façons et il semble difficile de l'inclure uniquement comme attribut d'une classe dans le diagramme « cycle parasitaire général » (par exemple en attribut d'une classe « entrée dans l'organisme hôte »). En effet, soit la liste des modes de transmission serait exhaustive et donc probablement illisible, soit elle serait réduite, et donc très peu informative du processus réel.

Un premier diagramme a donc été construit dans le but de décrire la façon dont un organisme pathogène *rentre* dans un nouvel hôte.

Nous avons recherché les différents modes de transmission qui pouvaient exister autour tout d'abord de l'hôte humain. Nous nous sommes alors rendu compte que le mode de transmission ne pouvait être un objet simple, mais qu'il devait être décomposé. Les différents éléments de décomposition que nous avons créés et qui

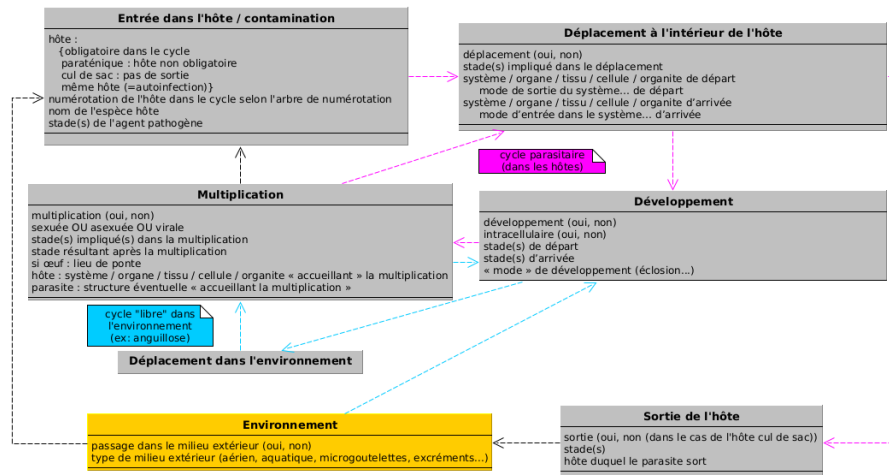


FIGURE 5. Diagramme UML générique décrivant le cycle parasitaire d'un agent pathogène, quel qu'il soit. Se succèdent: entrée dans l'hôte, déplacement dans l'hôte, développement, multiplication, sortie, passage dans l'environnement, développement, déplacement dans l'environnement, multiplication. Les flèches roses concernent les étapes intra-hôte, les flèches bleues les étapes dans l'environnement. En fonction des agents pathogènes, certaines étapes du cycle peuvent être absentes.

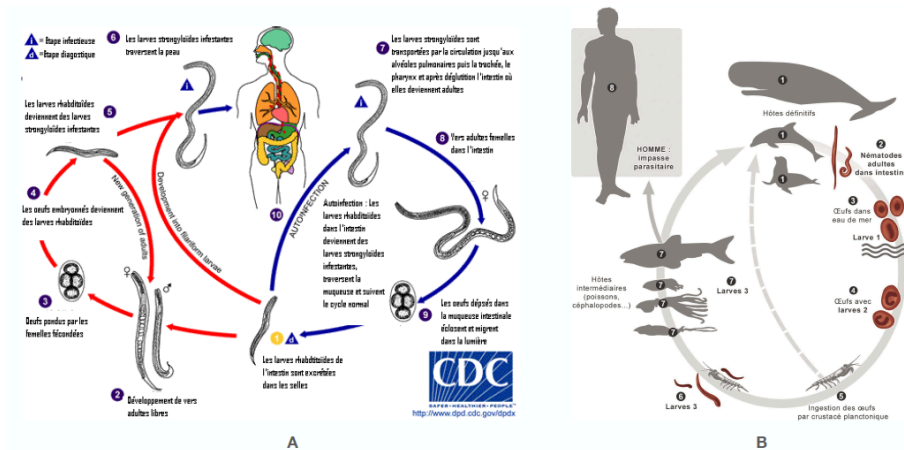


FIGURE 6. A Cycle Anguilliose , B Cycle Anisakis

caractérisent la transmission d'un agent pathogène depuis un hôte A vers un hôte B sont:

- la voie de contamination (ex: système musculaire, système digestif...)
- le mode de contamination (ex: transépithélial par passage traumatique via une lésion déjà existante, non transépithélial...)
- l'évènement permettant la contamination (ex: morsure, rapport sexuel, respiration...)

Une fois ces catégories créées et des exemples ajoutés, il est apparu que ces exemples étaient difficilement organisables, car ils regroupaient des interactions entre des hôtes invertébrés et vertébrés de toutes les façons possibles. Nous avons donc choisi de créer 4 diagrammes distincts pour décrire l'interaction de contamination :

- vertébré -> vertébré
- vertébré -> invertébré
- invertébré -> vertébré
- invertébré -> invertébré

Enfin, nous avons séparé un type de transmission qui aurait pu être inclus dans le diagramme « invertébré -> vertébré » ainsi que dans le diagramme « vertébré -> invertébré »: celle depuis un hôte vecteur invertébré vers un vertébré. En effet, cette relation est spécifique, redondante, et « simple à décrire », il nous est donc apparu judicieux de la séparer afin de faciliter la construction des diagrammes d'une part, et d'augmenter le confort des utilisateurs de la base de donnée d'autre part.

De la même façon que le diagramme « cycle parasitaire général » a induit la construction du diagramme « entrée dans les organismes », il a induit celle du diagramme « sortie des organismes ».

Nous avons également décomposé la sortie en trois éléments :

- la voie de sortie (système digestif, système urinaire...)
- le mode de sortie
- l'évènement permettant la sortie (toux et éternuement, contact depuis une muqueuse...)

### **3.3. Schématisation des processus**

La schématisation des diagrammes « entrée des organismes » et « sortie des organismes » nous a permis d'aborder l'idée de la schématisation générique de tout processus.

Les processus relèvent du fonctionnement des objets, ce qui implique une dynamique, une séquence temporelle entre différentes classes.

Dans les diagrammes déjà créés, d'autres processus ont été réifiés. Afin de simplifier la tâche de « transfert vers un schéma de base de données », nous avons cherché à généraliser les processus, de sorte à ce qu'ils puissent tous suivre le même modèle.

Pour ce faire, nous avons en parallèle :

- simplifié les diagrammes « entrée des organismes » et « sortie des organismes »,
- analysé les processus des diagrammes déjà construits,
- extrait de ces deux « travaux » les éléments les plus généraux permettant de décrire un processus.

*Résultat :*

La réflexion autour des processus, nous a amené à proposer une généralisation sous forme de chaîne de traitements ou *workflow*. Inspirés par les propositions de (Lin, 2011) nous proposons le diagramme suivant ( Figure 7).

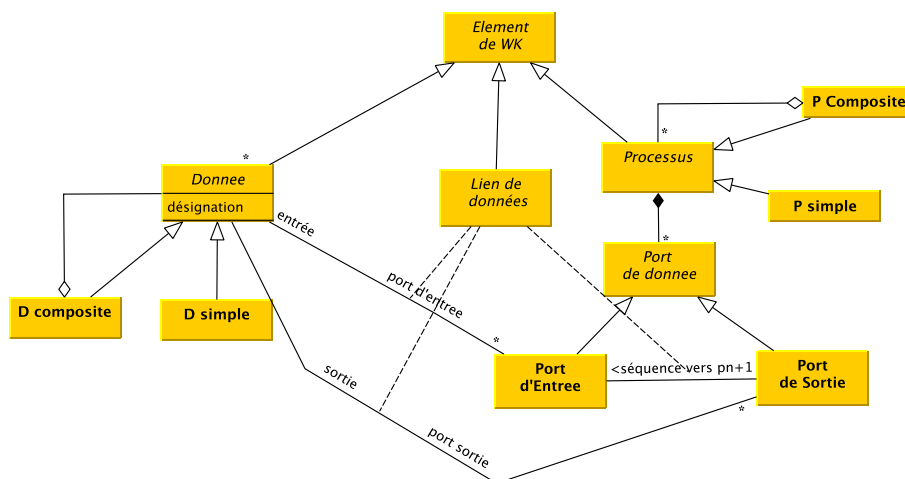


FIGURE 7. *Modèle de Workflow*

Les éléments d'un *workflow* sont des données liées à des processus par des liens. Tout processus simple ou composite comporte des ports (port d'entrée ou de sortie). Les liens de données peuvent être contraints selon la nature du processus.

Nous avons ensuite testé ce modèle sur le processus le plus compliqué déjà construit : le « cycle parasitaire général »

Le diagramme Figure 8 reprend sous la forme de processus uniquement (représentés avec leurs port d'entrée et de sortie) le diagramme « cycle parasitaire général » Figure 5

Il contient les diagramme « entrée des organismes » et « sortie des organismes » précédemment effectué et lui associe 6 autres processus issus du diagramme « cycle

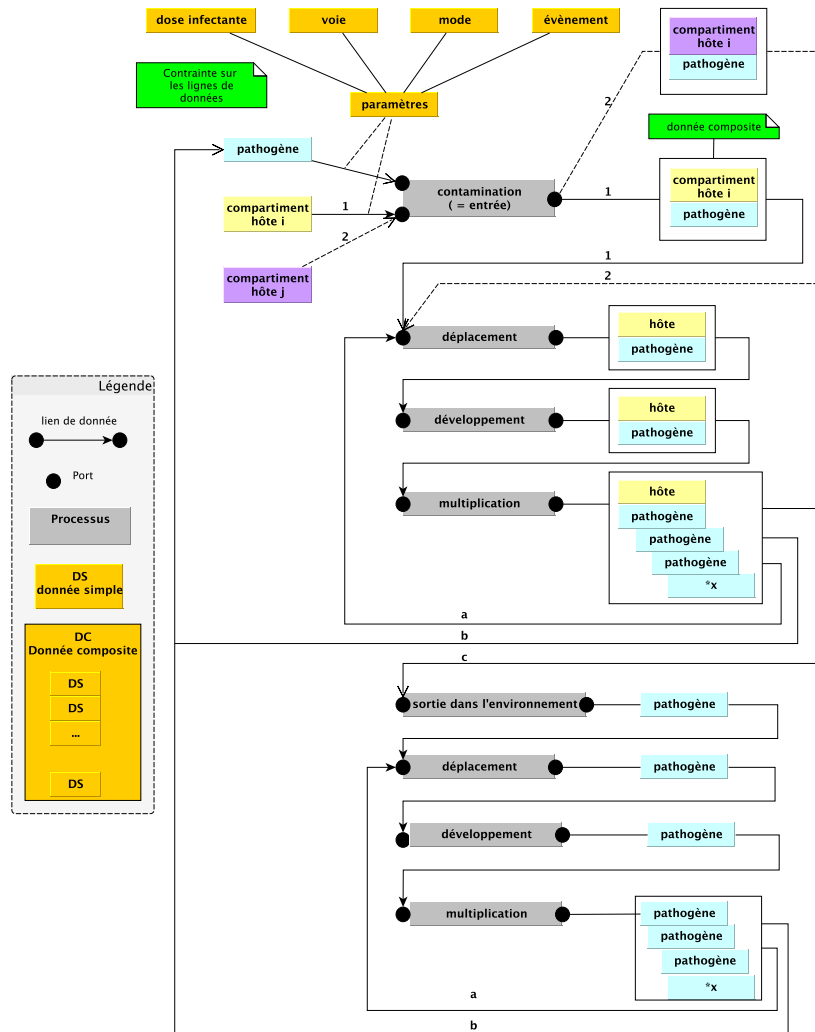


FIGURE 8. Cycle de contamination général

parasitaire général », reliés entre eux pour former une boucle : déplacement, développement, multiplication, sortie dans l'environnement, développement, multiplication.

#### 4. Spécifications de la plateforme

La plateforme doit :

- Assurer l'intégration des données issues des descriptions des divers diagrammes conceptuels au sein d'une base de données ;

– Assurer que toute donnée sauvegardée est référencée par une métadonnée liée à la référence bibliographique dont elle a été extraite. De plus si la connaissance sur la donnée est mise à jour, la métadonnée doit aussi enregistrer celle-ci. Le schéma conceptuel relatif aux métadonnées a été co-construit (cf. Figure 9)

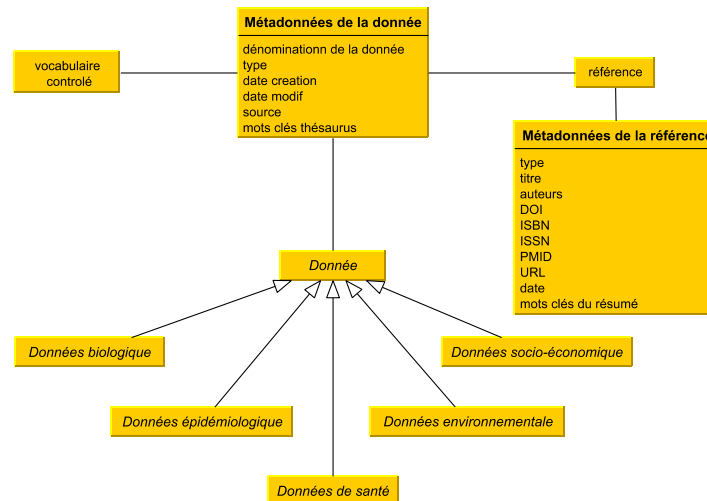


FIGURE 9. Les métadonnées

– Proposer une interface de recherche pour les utilisateurs soucieux d’acquérir des connaissances. Les recherches vont naturellement reposer comme point de départ sur les métadonnées et les vocabulaires partagés que celles-ci ont invoqué avant d’interroger le contenu de la base.

Du point de vue opérationnel cela implique d’une part la construction de la base de données comprenant les métadonnées liant les références bibliographiques et les concepts (structure et processus) et d’autre part la conception de l’interface d’accès proposée aux futurs utilisateurs soit spécialistes, soit grand public.

Pour un tel projet où la connaissance évolue sans cesse, la complexité et la variété des données sont importantes, le choix de la plateforme impose une réflexion approfondie sur l’architecture et les systèmes retenus (SGBD relationnels et/ou NoSql, thesaurus, ontologie, …).

## 5. Conclusion

Le projet nous a permis de mesurer l’importance de la pluridisciplinarité et celle de l’interdisciplinarité. Il était clair dès le départ que plusieurs disciplines allaient forcément intervenir et travailler dans le projet, cependant elles n’étaient pas forcément en interaction. La complexité des phénomènes étudiés nous invite en effet à ne plus disjoindre les savoirs ni les disciplines scientifiques et à questionner les objets

d'étude avec un œil extra-disciplinaire. (Morin, 1986), (Morin, 2014). Ce genre de projet, permet de développer une stratégie de résolution des confrontations voire des conflits grâce à des transferts de méthodes, de concepts et d'outils. Nous touchons alors à la richesse et à l'innovation de l'interdisciplinarité. Certes la mise en œuvre de la démarche reste simple (cf. Figure 2), manuelle mais pour l'instant elle a permis, à notre petit groupe, de partager, de co-construire et de naviguer entre généralisation et spécialisation de concepts.

Les perspectives sont évidemment liées à la description des domaines de données complémentaires. La formalisation des vocabulaires partagés utilisés pour les méta-données est d'ores et déjà une étape complémentaire en cours effectuée à partir de la terminologie (évoquée en section 2) et avec recherche parmi les thesaurus et ontologies existants. L'usage des traitements informatiques liés à la terminologie sera aussi envisagé pour automatiser les propositions de termes extraits des sources bibliographiques désignant les objets des diagrammes (modèles UML). Restera à préciser les solutions nécessaires à la résolution des demandes des usagers de la plateforme (requêtes sur métadonnées et sur données).

## Bibliographie

- Bates P. A. (2007). Transmission of leishmania metacyclic promastigotes by phlebotomine sand flies. *International journal for parasitology*, vol. 37, n° 10, p. 1097–1106. Consulté sur <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2675784/>
- Bizouarn P. (2016). L'éco-épidémiologie-vers une épidémiologie de la complexité. *médecine/sciences*, vol. 32, n° 5, p. 500–505.
- Blanchet C., Collin O., Boudet M., Delmotte S., Gilquin H., Guillaume J.-F. et al. (2019). Ifb-biosphère: Services cloud pour l'analyse des données des sciences de la vie. In *Journées réseaux-jres 2019*.
- Clément P. (2004). Science et idéologie: exemples en didactique et épistémologie de la biologie. In *Actes du colloque sciences, médias et société. ens-lsh*, p. 53–69.
- Combes C., Gavotte L., Moulia C., Sicard M. (2018). *Parasitisme, écologie et évolution des interactions durables*. Dunod.
- Cycle évolutif de l'anisakiose*. (2016). Consulté sur <http://campus.cerimes.fr/parasitologie/enseignement/toxocarose/site/html/2.html>
- Evans B., Leighton F. et al. (2014). A history of one health. *Rev Sci Tech*, vol. 33, n° 2, p. 413–420.
- Goehrs J.-M., Borel T., Costagliola D. (2012). La mise en place des cohortes en france: pourquoi, pour qui, comment et avec quels moyens? *Thérapie*, vol. 67, n° 4, p. 375–380.
- Kembellec G. (2019). Produire, analyser et partager des données ouvertes en humanités numériques: quelques bonnes pratiques. In *12ème colloque international d'isko-france: Données et mégadonnées ouvertes en shs: de nouveaux enjeux pour l'état et l'organisation des connaissances?*
- Lin Y. (2011). *Méthodologie et composants pour la mise en oeuvre de workflows scientifiques*. Thèse de doctorat non publiée, Montpellier 2.

- Loftus B., Anderson I., Davies R. *et al.* (2005). The genome of the protist parasite *Entamoeba histolytica*. *Nature*, vol. 433, p. 865–868.
- Morin E. (1986). *La connaissance de la connaissance* (vol. 3). Seuil.
- Morin E. (2014). *La tête bien faite. repenser la réforme, réformer la pensée*. Média Diffusion.
- Muller P.-A., Gaertner N. (2000). *Modélisation objet avec uml* (vol. 514). Eyrolles Paris.
- Pereira A., Cavalcanti N., Nascimento G. *et al.* (2013). Morphological and morphometric study of cercariae and adult worms of *Schistosoma mansoni* (slm strain) isolated from infected mice. Consulté sur <https://link.springer.com/article/10.1007/s00436-012-3235-9#citeas>
- Plateforme E. Pour la, SPF P. A. (s. d.). Covid-19 et faune sauvage.
- Strongyloidiasis biology*. (2019). Consulté sur <https://www.cdc.gov/parasites/strongyloides/biology.html>
- Zinsstag J., Schelling E., Waltner-Toews D., Whittaker M. A., Tanner M. (2020). *One health, une seule santé: Théorie et pratique des approches intégrées de la santé*. éditions Quae.



---

## **ParkinsonCom : Outil d'Aide à la Communication pour Personnes atteintes de la Maladie de Parkinson**

**Káthia Marçal de Oliveira<sup>1</sup>, Nejmeddine Allouche<sup>1</sup>,  
Véronique Delcroix<sup>1</sup>, Yohan Guerrier<sup>1</sup>, Christophe Kolski<sup>1</sup>,  
Sophie Lepreux<sup>1</sup>, Philippe Pudlo<sup>1</sup>, Yosra Rekik<sup>1</sup>, Elise Batselé<sup>2</sup>,  
Mathilde Boutiflat<sup>2</sup>, Mark Freyens<sup>2</sup>, Hélène Geurts<sup>2</sup>,  
Romina Rinaldi<sup>2</sup>, Loïc Dehon<sup>3</sup>, Nicolas Jura<sup>3</sup>**

1. Univ. Polytechnique Hauts-de-France, LAMIH CNRS, UMR 8201  
Le Mont Houy, cedex 9, 59313, Valenciennes, France  
[prenom.nom@uphf.fr](mailto:prenom.nom@uphf.fr)

2. Université de Mons  
Place du Parc, 20 - 7000 Mons, Belgique  
[prenom.nom@umons.ac.be](mailto:prenom.nom@umons.ac.be)

3. Drag'On Slide SRL  
Mons, Belgique  
[info@dragonslide.com](mailto:info@dragonslide.com)

---

*RESUME. La maladie de Parkinson est la deuxième maladie neurodégénérative la plus répandue dans le monde. Les personnes atteintes signalent des troubles de la communication qui détériorent l'intelligibilité de leur discours, leur capacité à exprimer leurs états affectifs et, par conséquent, leurs relations sociales. Dans ce contexte, nous travaillons au développement d'un outil d'aide à la communication pour les personnes atteintes de la maladie de Parkinson afin d'améliorer leur participation et leur inclusion sociale. Cet article présente le premier prototype développé et évalué par des personnes atteintes de la maladie de Parkinson.*

*ABSTRACT. Parkinson's disease is the second most common neurodegenerative disease in the world. People with the disease report communication disorders that impair the intelligibility of their speech, their ability to express their emotional states and, consequently, their social relationships. In this context, we have been working on the development of a communication tool for people with Parkinson's disease to improve their participation and social inclusion. This article presents the first prototype developed and evaluated by people with Parkinson's.*

*Mots-clés : conception centrée utilisateur, parkinson, communication.*

*KEYWORDS: user-centered design, parkinson, communication*

---

## 1. Introduction

La maladie de Parkinson (MP) est la deuxième maladie neurodégénérative la plus répandue après la maladie d'Alzheimer. Les répercussions de cette maladie sur la qualité de vie et la participation sociale en font une préoccupation majeure en termes de santé publique. Les symptômes de la MP sont classiquement divisés en symptômes moteurs (les plus connus) et en symptômes non moteurs (Sveinbjornsdottir, 2016). Parmi les symptômes non-moteurs, 70% des personnes atteintes de la MP font état de troubles de la communication au niveau de la parole et de la voix. Ces déficiences entraînent le plus souvent un retrait social actif ou passif et ont un impact négatif sur le bien-être subjectif des patients.

Sur la base de ces observations, nous avons lancé un projet Interreg, nommé ParkinsonCom (<https://parkinsoncom.eu>) (Oliveira et al., 2021), qui vise à développer un outil d'aide à la communication pour les personnes atteintes de la MP afin d'améliorer leur participation et leur inclusion sociale. Pour atteindre cet objectif, nous avons utilisé une approche de conception centrée utilisateur (Lepreux et al., 2021) de manière à prendre en compte les besoins réels de communication des personnes atteintes de la MP. Cette approche nous conduit à relever 2 principaux défis : (1) la conception d'un système simple d'utilisation via son interface utilisateur pour ce public, (2) l'adaptation du système en tenant compte du profil évolutif des personnes atteintes de la MP (compte tenu du caractère dégénératif de la maladie) (Guerrier et al., 2021). A partir d'une première itération de cette approche, nous avons développé un prototype qui a été évalué par des personnes atteintes de la MP.

Dans les sections suivantes, nous présentons la conception et l'évaluation de ce prototype, puis nous concluons l'article en présentant les travaux en cours.

## 2. Conception et Évaluation du prototype ParkinsonCom

En suivant le processus centré utilisateur basé sur la norme NF EN ISO 9241-210 (2011), nous avons commencé par la réalisation d'un *survey* pour identifier les principales difficultés de communication des personnes atteintes de la MP. Ensuite, nous avons mené des entretiens auprès d'experts du vécu (onze personnes atteintes de la MP, sept conjoints de personnes présentant la MP et trois neurologues). Suite à l'extraction des besoins prioritaires des utilisateurs, et en interaction avec les neurologues, deux fonctionnalités principales ont été définies :

- Préparer un dialogue - cette fonctionnalité serait utilisée pour les personnes atteintes de la MP pour communiquer avec leur entourage. Elles pourront préparer un dialogue en tapant du texte, ou en utilisant des pictogrammes que l'application proposera au fur et à mesure de leurs saisies. Ces dialogues pourront être lus immédiatement ou après par la synthèse vocale, ou enregistrés.
- Faire une demande d'aide à une personne spécifique en l'appelant, ou à une personne à proximité - cette fonctionnalité est utile en situation de blocage, connue comme *mode off* (Whitfield et Goberman, 2018). De plus, pour permettre aux personnes avec la MP d'accéder à un moment de détente en

attendant que cette période de blocage soit passée, un lecteur de musique ainsi qu'un lecteur d'histoires drôles sont également disponibles.

Les Figures 1 et 2 montrent les diagrammes d'activité et les interfaces utilisateur de ces deux fonctionnalités. En cas de situation de blocage, l'utilisateur peut sélectionner l'icône (pouce vers le bas) dans l'interface utilisateur « **Préparer un Dialogue** » (Error! Reference source not found. à gauche) pour accéder à la fonctionnalité de « **Faire une demande** » (Error! Reference source not found. à droite). Suivant l'approche centrée utilisateur, chaque élément de l'interface utilisateur a été choisi en fonction de l'avis de personnes atteintes de la maladie de Parkinson auxquelles différentes maquettes ont été présentées.

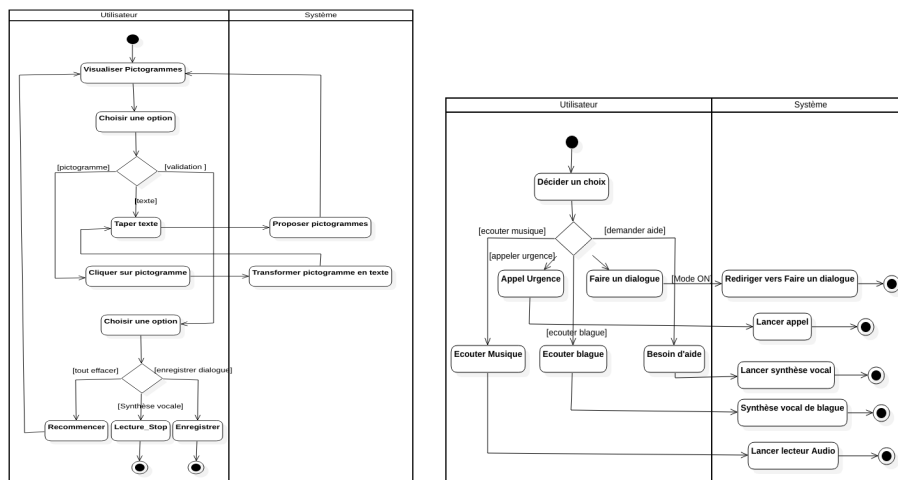


Figure 1. Diagrammes d'activités : “Préparer un dialogue” (à gauche) et “Faire une demande” (à droite)

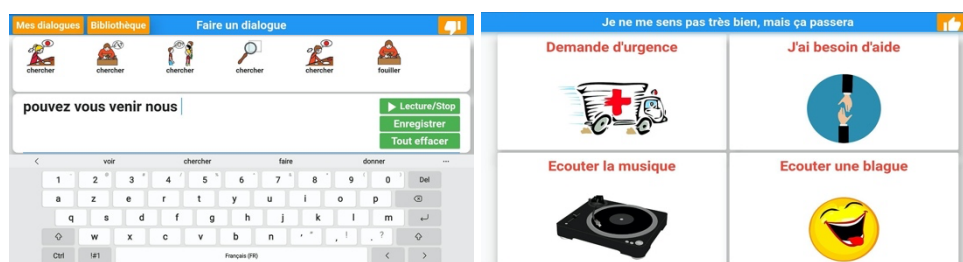


Figure 2. Interface utilisateur : “Préparer un dialogue” (à gauche) et “Faire une demande” (à droite)

Une première évaluation a été réalisée avec 26 personnes asymptomatiques pour évaluer l'utilisabilité en général. Ensuite, 16 personnes atteintes de la MP, en France

et en Belgique, ont évalué le prototype. L'ensemble des retours a permis de définir des améliorations relatives à des questions ergonomiques, fonctionnelles et en lien avec l'adaptation de l'interface utilisateur en tenant compte de l'évolutivité de la MP.

### 3. Travaux en cours

À partir de la conception et de l'évaluation initiales du prototype destiné à favoriser la communication entre les personnes atteintes de la MP et leur entourage, un ensemble d'exigences a été établi pour le développement de la version finale. Les travaux en cours visent à développer la version finale, puis à l'évaluer, dans le cadre de la démarche centrée utilisateur mise en place. En parallèle de ce développement, nous étudions l'utilisation de réseaux bayésiens pour adapter l'interface au profil de l'utilisateur en tenant compte des caractéristiques personnelles qui évoluent avec la maladie (comme la vision, la précision et la mémoire).

#### *Remerciements*

*Ce travail a été réalisé avec le soutien du Fonds Européen de Développement Régional Interreg et l'AVIQ (l'Agence pour une Vie de Qualité), Belgique.*

### Bibliographie

- NF EN ISO 9241-210 (2011) Comité Européen de Normalisation. 2010. Ergonomie de l'interaction homme-système–Partie 210: Conception centrée sur l'opérateur humain pour les systèmes interactifs.
- Sveinbjornsdottir, S. (2016) The clinical symptoms of Parkinson's disease. *Journal of neurochemistry*, 139, 318-324.
- Guerrier, Y., Oliveira, K., Kolski, C., Lepreux, S., Apedo, K., Delcroix, V., Ezzedine, H. (2021). Une étude systématique pour la conception d'un système d'aide à la communication pour les personnes atteintes de la Maladie de Parkinson. Actes du 39ème Congrès INFORSID, Dijon, France, pp. 71-86, juin.
- Lepreux, S., Apedo, K., Oliveira, K. (2021). Vers une conception centrée sur l'utilisateur ayant un profil évolutif : Une étude de cas avec des personnes atteintes de la maladie de Parkinson. *IHM'21, Proceedings of the 31st Conference on the Interaction Homme-Machine: Adjunct*, ACM, Metz, France, avril.
- Oliveira K., Batselé E., Lepreux S., Buchet E., Kolski C., Boutiflat M., Delcroix V., Geurts H., Apedo K., Dehon L., Ezzedine H., Guerrier Y., Haelewyck M., Jura N., Pudlo P., Rekik, Y. (2021). ParkinsonCom Project: Towards a Software Communication Tool for People with Parkinson's Disease. In *International Conference on Human-Computer Interaction* (pp. 418-428). Springer, Cham.
- Whitfield J., Reif A., Goberman A. (2018). Voicing contrast of stop consonant production in the speech of individuals with Parkinson disease ON and OFF dopaminergic medication. *Clinical linguistics & phonetics*, 32(7), 587-594.