43ÈME ÉDITION

INFORSID 2025

Éthique, équité et systèmes d'information

Construire un numérique responsable et inclusif

iNFORSID

Du 3 au 6 juin 2025

UNIVERSITÉ DE PAU
ET DES PAYS DE L'ADOUR

Présidente du comité de programme : Elsa Negre

Présidente du forum JCJC : Lylia Abrouk

Présidents du comité d'organisation :
Christian Sallaberry - Philippe Roose

UNIVERSITÉ DE PAU ET DES PAYS DE L'ADOUR
COLLÈGE STEE | EA 3000 LIUPPA

HUPI | PAU BÉARN PYRÉNÉES Communauté d'Agglomération | CGI

inforsid-contact@univ-pau.fr

+ d'infos et programme
https://inforsid2025.sciencesconf.org

Préface

Le 43e congrès INFORSID, qui se déroule cette année à Pau, au sein de l'Université de Pau et des Pays de l'Adour, se veut bien plus qu'un simple événement scientifique. Il incarne un moment crucial pour réfléchir ensemble aux transformations profondes que le numérique imprime sur nos sociétés. En 2025, alors que les Systèmes d'Information (SI) sont au cœur des enjeux sociétaux, la question de leur impact éthique, social et environnemental n'a jamais été aussi urgente.

Les organisations et le monde vivent actuellement de grandes transformations, largement liées aux technologies de l'information et à leurs impacts. Ces transformations, outre leur aspect technique et organisationnel, imposent de nouvelles responsabilités en matière d'éthique, d'équité et de justice sociale, afin que les SI soutiennent une société plus inclusive et équitable. La rapidité des changements dans les organisations, ainsi que les défis sociétaux et environnementaux, nécessitent de mettre en place des processus allant au-delà de l'amélioration continue et d'envisager des transformations plus fondamentales. Dans ce contexte, l'innovation, la créativité, mais aussi la responsabilité, sont des facteurs déterminants. L'imprévisibilité de ces mêmes transformations (notamment les effets indirects pervers) exige plus que jamais une vision systémique dans l'ingénierie et la gouvernance des SI.

Le thème de cette édition, « **Éthique, Équité et Systèmes d'Information : construire un numérique responsable et inclusif** », propose de repenser la manière dont les SI peuvent être des moteurs de changement positif. À une époque où la technologie modifie les règles de nombreux secteurs – de la santé à l'éducation, de la finance à l'écologie –, il devient essentiel de remettre l'humain au centre de cette révolution numérique. L'objectif n'est pas seulement de développer des systèmes plus performants, mais de les rendre responsables, inclusifs et respectueux des principes d'éthique et d'équité.

Les SI, à la fois techniques et scientifiques, offrent d'innombrables opportunités d'évolution et de transformation. Ces opportunités supposent une capacité à capter, stocker, organiser, rechercher, analyser et apprendre à partir de gros volumes d'informations. De nouveaux problèmes inédits émergent. Dans ce contexte, il est primordial d'être pleinement conscients des défis auxquels nous faisons face dans le nouveau monde VUCA (*Volatility, Uncertainty, Complexity, Ambiguity*). Nous, en tant qu'ingénieurs, chercheurs, enseignants et citoyens, sommes responsables de l'impact de ces technologies.

Le programme du congrès a été conçu pour interroger ces enjeux majeurs. En invitant des experts internationaux, comme :

- *SERGIO ILARRI* (LSI, Universidad de Zaragoza), expert en SI dans les organisations : « ***Gestion des données pour une société plus durable et responsable*** »[1], *et*
- *OLIVIA TAMBOU* (Cr2D, Université Paris-Dauphine), juriste et spécialiste de l'éthique de l'Intelligence Artificielle (IA) : « ***L'approche européenne de l'encadrement juridique et éthique de l'IA*** »,

ce congrès permet de porter un regard critique sur la manière dont les SI transforment nos sociétés. Leur intervention met en lumière les défis juridiques, éthiques et pratiques d'une régulation de plus en plus nécessaire dans un monde où l'IA et les nouvelles technologies redéfinissent les règles du jeu.

Les trois ateliers de cette année :

- « **Accompagner les acteurs de la transition écoresponsable des SI** » porté par *ADEL NOUREDDINE ET NATHALIE VALLES-PARLANGEAU*,
- « **Aider à concevoir des SI socialement responsables** » porté par *MARYSE SALLES, LYCETTE CORBION, GABRIEL COLLETIS, PHILIPPE NEUVILLE ET NATHALIE VALLES-PARLANGEAU*,
- « **Détection d'anomalies dans l'environnement** » porté par *LYLIA ABROUK ET ALEXIS GUYOT*,

offrent également des espaces d'échanges concrets, où chaque participant peut réfléchir à l'intégration de principes éthiques dans les projets de SI, à la conception de systèmes écoresponsables, ou encore aux moyens de renforcer la gouvernance numérique. Des moments d'interactivité comme ceux-ci permettent de dépasser le cadre académique pour plonger dans des pratiques réellement applicables.

Au cœur de cette édition se trouvent aussi les jeunes chercheurs, 18 participants en particulier, qui, à travers le forum Jeunes Chercheuses et Chercheurs (JCJC), ont l'opportunité de présenter leurs travaux innovants et de confronter leurs idées avec des experts. Enfin, la remise du Prix de thèse INFORSID 2025 : « Co-conception d'un produit et de son système industriel : Une approche ingénierie des exigences pour l'aérospatial », vient récompenser Mme CHAN Anouck (ISAE SupAero, ONERA) pour ses contributions au domaine des SI.

Le déploiement des SI dans cet environnement complexe, avec des données sensibles/massives/hétérogènes, génère des risques juridiques, sociaux et financiers, rendant la sécurité, et notamment la cybersécurité, centrale dans les préoccupations des directions des systèmes d'information (DSI), des chercheurs et des équipes pédagogiques en SI. Face à ce déploiement de solutions technologiques gourmandes en

[1] Ce travail fait partie du projet PID2020-113037RB-I00, financé par MICIU/ AEI/10.13039/501100011033.

ressources énergétiques, les SI doivent aussi répondre aux défis environnementaux pour proposer une informatique durable et un numérique responsable.

Concernant l'organisation scientifique du congrès en elle-même, nous avons cette année reçu 28 soumissions :

- 17 articles pour la catégorie « Articles originaux (longs et courts) »
- 11 articles pour la catégorie « Articles déjà publiés à l'international »[2]

Les articles ont fait l'objet d'une évaluation par les membres du comité de programme (CP). À l'issue de cette évaluation, les rapporteurs ont confronté leurs analyses afin de converger vers une décision partagée, laquelle a été discutée et finalisée lors de la réunion de sélection des articles. Tout ce travail d'évaluation et d'échanges est un travail très important qui plus est parfois réalisé dans des délais relativement courts. Je tiens à adresser un grand merci à l'ensemble des membres du CP, dont l'implication rigoureuse et bienveillante a été essentielle. Grâce à leur travail de relecture approfondie et, pour certains, à leur accompagnement personnalisé des auteurs, cette édition a pu voir le jour dans les meilleures conditions.

Suite à la réunion de sélection des articles, nous avons retenu pour cette année :

- 10 articles pour la catégorie « Articles originaux » (10 articles longs)
- 11 articles pour la catégorie « Articles déjà publiés à l'international »

Sur la base des articles retenus, le programme du congrès est construit et organisé autour des six sessions suivantes :

- Ethique, Responsabilité & Environnement
- Représentation des connaissances
- Ethique, Equité & Connaissance
- Transformation numérique & Processus métier
- IA & Comportement humain
- Données & Gouvernance

L'organisation de cet événement n'aurait pu se faire sans l'implication précieuse de *Christian Sallaberry* et *Philippe Roose*, qui ont codirigé ce congrès avec professionnalisme et une vision commune. Leur coordination sans faille et leur attention aux détails ont été des éléments déterminants pour assurer la réussite de cette édition.

Le travail des membres du comité scientifique, des responsables d'ateliers, des conférenciers invités *Olivia Tambou* et *Sergio Ilarri,* et celui de *Lylia Abrouk*, responsable du forum JCJC, a été essentiel pour garantir la qualité des échanges et des discussions. Leur engagement à faire de ce congrès un lieu de réflexion, d'apprentissage et

[2] Articles déjà publiés en 2024 dans d'excellentes conférences ou revues internationales, soumis pour présentation avec une synthèse en français (2 pages max).

d'inspiration est un atout inestimable pour la communauté des SI et INFORSID en particulier.

J'adresse un petit clin d'œil aux membres du bureau de l'association INFORSID qui m'ont fait l'honneur de me confier la tâche de Présidente du Comité de Programme. Je les remercie vivement pour la confiance qu'ils m'ont témoignée et l'aide continue qu'ils m'ont apportée tout au long de ces derniers mois.

Enfin, un remerciement tout particulier aux organisateurs locaux et à travers eux, à l'ensemble des soutiens (institutionnels ou industriels), qui, grâce à leur accueil chaleureux et à leur efficacité, ont permis de faire de ce congrès un véritable succès. Leur travail en amont et pendant l'événement a créé un environnement propice à la collaboration et à l'échange.

À toutes et à tous les participants, un grand merci pour votre présence et vos contributions. Ces actes témoignent de l'importance des échanges que nous avons partagés à Pau, et j'espère que les discussions ici entamées nourriront de nouvelles réflexions et actions autour de l'éthique, de l'inclusion et de la responsabilité dans les SI.

Merci à toutes et à tous !

Elsa NEGRE

Présidente du comité de Programme d'INFORSID 2025

Membres du Comité d'Organisation (CO)

Co-présidents :

- CHRISTIAN SALLABERRY, LIUPPA, Université de Pau et des pays de l'Adour (Pau)
- PHILIPPE ROOSE, LIUPPA, Université de Pau et des pays de l'Adour (Bayonne)

Membres :

- MARIE-NOËLLE BESSAGNET
- CHARLOTTE DARRICADES
- MOHAMMED ERRITALI
- NICOLAS EVAIN
- ERNESTO EXPOSITO GARCIA
- SEBASTIEN LABORIE
- ANNIG LE PARC LACAYRELLE
- ZHONGWEI MA
- CHRISTOPHE MARQUESUZAÀ
- MAXIME MASSON
- THIERRY NODENOT
- ADEL NOUREDDINE
- NICOLAS TIREL
- NATHALIE VALLÈS-PARLANGEAU
- YUNJI ZHANG

Membres du Comité de Programme (CP)

Lylia Abrouk, LIB - Université de Bourgogne

Pascal Andre, LS2N - Université de Nantes

Bernd Amann, Sorbonne Université

Pierre-Emmanuel Arduin, DRM - Université Paris-Dauphine

Ladjel Bellatreche, LIAS/ENSMA - Université de Poitiers

Nourhene Ben Rabah, CRI - Université Paris 1 Panthéon-Sorbonne

Sandra Bringay, Université de Montpellier

Armelle Brun, LORIA - Université de Lorraine

Guillaume Cabanac, IRIT - Université de Toulouse

Sylvain Castagnos, LORIA - Université de Lorraine

Alexandre Chanson, LIFAT - Université de Tours

François Charoy, LORIA - Université de Lorraine

Max Chevalier, IRIT - Université de Toulouse

Camelia Constantin, LIP6 - Université Paris 6

Mario Cortes-Cornax, LIG - Université Grenoble Alpes

Nadine Cullot, Université de Bourgogne

Rebecca Deneckere, CRI - Université Paris 1 Panthéon-Sorbonne

Cedric Du Mouza, Cédric - CNAM

Cyril Faucher, L3i - La Rochelle Université

Cecile Favre, ERIC - Université Lyon 2

Faiza Ghozzi, MIRACL - Université de Sfax

David Gross-Amblard, Université de Rennes

Nicolas Herbaut, CRI - Université Paris 1 Panthéon-Sorbonne

Lydia Khelifa Chibout, Centre Scientifique et Technique du Bâtiment

MANUELE KIRSCH PINHEIRO, Université Paris 1 Panthéon-Sorbonne

ELENA KORNYSHOVA, Cédric - CNAM

CHRISTINE LAHOUD, CIAD – UTBM

PIERRE LARMANDE, IRD

ANNE LAURENT, LIRMM - Université de Montpellier

SABINE LOUDCHER, ERIC - Université Lyon 2

SOFIAN MAABOUT, LaBRI - Université de Bordeaux

MAUDE MANOUVRIER, LAMSADE - Université Paris-Dauphine

IMEN MEGDICHE, IRIT - INU Champollion

KATHIA OLIVEIRA, LAMIH - Université Polytechnique Hauts-de-France

PASCAL PONCELET, Université de Montpellier

MATHIEU ROCHE, CIRAD

CATHERINE ROUSSEY, INRAE

PHILIPPE ROOSE, LIUPPA - Université de Pau et des Pays de l'Adour

INES SAAD, MIS - Université Jules Verne Picardie

CHRISTIAN SALLABERRY, LIUPPA - Université de Pau et des Pays de l'Adour

MARINETTE SAVONNET, LE2I - Université de Bourgogne

DIDIER SCHWAB, Université de Grenoble

SANA SELLAMI, Aix Marseille Université

JIEFU SONG, IRIT - Université Toulouse Capitole

RONAN TOURNIER, IRIT - Université Toulouse Capitole

NATHALIE VALLES, LIUPPA - Université de Pau et des Pays de l'Adour

MARLENE VILLANOVA, LIG - Université Grenoble Alpes

Programme synthétique INFORSID 2025

INFORSID 2025

3-6 Juin 2025 - Université de Pau et des Pays de l'Adour

SEMAINE DU : 02/06/2025	MARDI 3	MERCREDI 4	JEUDI 5	VENDREDI 6
8h	Enregistrement / Accueil	Enregistrement / Accueil	Enregistrement / Accueil	Enregistrement / Accueil
8h30		OUVERTURE	Keynote 2 : Olivia Tambou (Cr2D, Université Paris-Dauphine)	Session 5 IA & Comportement humain (3 articles) / Session 6 Données & Gouvernance (3 articles)
9h	Atelier : *Accompagner les acteurs de la transition éco-responsable des SI*	Keynote 1 : Ilarri Sergio (LSI, Universidad de Zaragoza)		
9h30			*Pause café*	
10h	*Pause café*	*Pause café*	Forum JCJC	*Pause café*
10h30		Session 1 Ethique, Responsabilité & Environnement (4 articles) / Session 2 Représentation des connaissances (3 articles)		Prix de thèse + Prix stage SD
11h				
11h30				CLOTURE
12h				
12h30	*Déjeuner*	*Déjeuner*	*Réunion CE* / *Déjeuner*	
13h				
13h30				
14h	Atelier : *Aider à concevoir des SI socialement responsables* / Atelier : *DAE*	Table ronde	Session 3 Ethique, Equité & Connaissance (4 articles) / Session 4 Transformation numérique & Processus métier (4 articles)	
14h30				
15h				
15h30		*Pause café*		
16h	*Pause café*		*Pause café*	
16h30				
17h		Moment convivial	Assemblée Générale INFORSID	
17h30				
18h			Moment convivial	
18h30				
19h				
19h30				
20h			Soirée de Gala	
20h30				

Prix de thèse 2025 de l'association INFORSID

L'association INFORSID félicite *MME CHAN ANOUCK* pour sa thèse soutenue au cours de l'année 2024 et élue ***Prix de Thèse 2025***.

Titre de la thèse : Co-conception d'un produit et de son système industriel : Une approche ingénierie des exigences pour l'aérospatial

Laboratoire : ISAE SupAero, ONERA

Directeurs de thèse : Thomas Polacsek, Stéphanie Roussel

Date de soutenance : 13/11/2024

Mots-clefs : Ingénierie des exigences, Co-conception, Modélisation par objectifs, Systèmes industriels aérospatiaux, Collaboration multi-acteurs.

Table des matières des actes

« Conférences invitées » - Animation : Elsa Negre, Philippe Roose, Christian Sallaberry

« Ethique, Responsabilité & Environnement » - Animation : Elena Kornyshova

« Représentation des connaissances » - Animation : Cyril Faucher

« Ethique, Equité & Connaissance » - Animation : Pierre-Emmanuel Arduin

« Transformation numérique & Processus métier » - Animation : Rebecca Deneckere

« IA & Comportement humain » - Animation : Max Chevalier

« Données & Gouvernance » - Animation : Franck Ravat

Légende

Internat. **Papier International**

Long **Papier Long**

Data management for a more sustainable and responsible society[1]

Professor Sergio ILARRI

I3A, Universidad de Zaragoza (Spain)
silarri@unizar.es

1. Résumé

Applying appropriate data engineering and data management practices is essential to achieve sustainability in a society. For example, it is a cornerstone and fundamental building block for AI applications. On the one hand, it is clear that suitable data services can contribute to achieving Sustainable Development Goals (SDGs), which are a political priority. This is related to the idea of green with software. On the other hand, applications and data services should be developed with an awareness of their sustainable impact. Therefore, the data management approaches themselves should be sustainable, according to the idea of green software. While there has been significant progress in terms of research into reducing energy consumption in computing (Green IT), very little work has been done in the design and development of data management proposals where the integration of sustainability is a fundamental characteristic, not only for energy saving but also for other types of sustainability, such as mobility sustainability.

By designing appropriate data management techniques, we can contribute to a better daily life for people and also to foster suitable behaviors in such a way that the existing modern challenges that we face as a society can be tackled. Thus, traditional data management techniques used to develop information services and applications for mobile users can be adapted to be useful, effective, and impactful in the current circumstances that permeate our lives, considering aspects such as the existing environmental risks. Besides, we have to take into account that the previously-existing challenges of Big Data have not been completely solved yet; thus, the existence of large amounts of heterogeneous data, usually coming at high speeds and in large quantities from a variety of data sources (sensors, other users, etc.), makes their collection, retrieval, filtering and processing challenging. Personalization, both as a means to deal with the existing data flood and to offer really customized information adapted to the needs and current context of each specific user, becomes of paramount importance to provide to each citizen the data that he/she really needs at the right moment.

[1] Research supporting this talk is part of the project PID2020-113037RBI00, funded by MICIU/AEI/10.13039/501100011033; a proposal of the new intended DAMASCOS project is currently under review (PID2024-157027OB-I00). Besides the NEAT-AMBIENCE project, we also thank the support of the Departamento de Ciencia, Universidad y Sociedad del Conocimiento del Gobierno de Aragón (Government of Aragon: Group Reference T64_23R, COSMOS research group).

Within the NEAT-AMBIENCE project, we tackle the design of novel data management techniques for mobile users and for drivers. On the one hand, we develop solutions for mobile Context-Aware Recommender Systems (CARS) which incorporate the consideration of novel factors such as ensuring a suitable physical distance among people (which we call Side-CARS —SocIal-Distance prEserving CARS—). On the other hand, we tackle challenges of developing information services for drivers, particularly services that provide information about resources on the roads that can be of interest to drivers (mobile resource search), such as parking spaces. Besides, in the context of NEAT-AMBIENCE we also explore specific use cases, related to tourism, resources for drivers, health and agriculture.

In this keynote, we will explore some of our initiatives in NEAT-AMBIENCE. Besides, we will also present some initial ideas about how data management can integrate sustainability as a key ingredient, not only from the perspective of energy efficiency, but also for other types of sustainability, such as organizational sustainability (i.e., more inclusive and egalitarian organizations) and mobility sustainability, focused not only on reducing the carbon footprint of transport systems, but also on enabling the circulation of citizens with a controlled impact, and where the poles of attraction of citizens (e.g., tourist destinations) are able to exploit their places of interest in a sustainable and culturally-attractive way. We should not forget either a third key axis, which is that of well-being sustainability, where data management in basic domains such as health or advances in more efficient food systems must also include the sustainability component. We intend to cover these aspects in detail in a future research project (DAMASCOS), whose main motivation is to progress in the development of data management techniques and services that integrate sustainability for the three aforementioned pillars: organizational, mobility and well-being sustainability. Sustainability plays a double role here: we aim at sustainable data management techniques to improve sustainability. Our intention is to advance the state of the art regarding the design and development of novel solutions for sustainability-oriented decision-making processes in several scenarios.

2. Biographie

Sergio Ilarri[2] is a Full Professor (*"Profesor Catedrático de Universidad"*; branch of knowledge: Engineering and Architecture) at the University of Zaragoza (Spain). He belongs to the Department of Computer Science and Systems Engineering and develops his work in the School of Engineering and Architecture of the University of Zaragoza. He is a Computer Science Engineer (University of Zaragoza, 2001) and Doctor in Computer Science (University of Zaragoza, 2006 — Doctoral Program in Systems and Informatics Engineering).

Currently, he is the principal investigator (PI) of the research group COS2MOS (Computer Science for Complex System Modeling, http://cos2mos.unizar.es; current reference: T64_23R), recognized as reference group (*"grupo de referencia"*) by the Government of Aragon for the period 2023-2025 (along with Ramón Hermoso as co-PI); he also played the role of PI of the group during the previous recognition period 2020-2022 (where the group was also recognized as reference group) and during the period 2017-2020 (during which the group had the previous category of group in development / *"grupo en desarrollo"*), thus leading the group to its consolidation. Besides, he is currently leading the national research project NEAT-AMBIENCE "Next-gEnerATion dAta Management to foster suitable Behaviors and the resilience of cItizens against modErN ChallEnges" (http://webdiis.unizar.es/~silarri/NEAT-AMBIENCE, PID2020-113037RBI00, funded by MICIU/AEI/10.13039/501100011033 —Spanish State Research Agency—) and has experience also as PI of previous research projects.

[2] http://webdiis.unizar.es/~silarri

His main research area is digital data management. His research interests include mobile computing, vehicular networks, context-aware recommender systems, data streams, information systems, data semantics, and data-driven decision making. His research on data management within the COSMOS group aligns with the development of data management techniques that allow addressing the challenges of the need to process and exploit large volumes of heterogeneous data in complex systems. Tasks such as the following are considered: development of data management techniques for highly-dynamic mobile computing environments (vehicle networks, mobile users in a city, etc.), and with support for peer-to-peer (P2P) mobile networks; design of strategies for the analysis of large volumes of data for the development of recommendation systems to help users, in particular applied to mobile computing and smart cities; and development of techniques and/or models for the exploitation of information to support decision-making in complex environments with uncertainty, where the information available may be incomplete or imprecise.

He has authored publications in relevant journals (more than 45 regular papers in JCR-indexed journals, with an average ranking position at the top 25%) and conferences and has participated as guest editor of special issues of journals such as Transportation Research Part C, IEEE Internet Computing, IEEE Multimedia, the Journal of Systems and Software, Distributed and Parallel Databases, and Personal and Ubiquitous Computing. For a year, he was a visiting researcher in the Mobile Computing Laboratory at the Department of Computer Science at the University of Illinois in Chicago, and he has also cooperated with other universities (e.g., through research stays with the University of Valenciennes and with IRIT in Toulouse —France—, as well as other short visits and joint actions). He has also participated in numerous program committees as well as in the organization of different events (e.g., Program Co-Chair of the 28th International Database Engineered Applications Symposium Conference —IDEAS 2024—, Program Co-Chair of the 11th International ACM Conference on Management of Digital EcoSystems —MEDES 2019—, Chair and local organizer of the 13th International Workshop on Semantic and Social Media Adaptation and Personalization —SMAP 2018—, Co-Chair of the Third International Workshop on Information Management in Mobile Applications —IMMoA 2013— in conjunction with VLDB 2013, track chair in several events, etc.),

Besides his teaching and research activities, he has served in several academic management positions (as director of university masters / *"Másteres Propios"*, subdirector of the Department of Computer Science and Systems Engineering at the University of Zaragoza from May 2016 to May 2020, and currently as the Coordinator of the Computer Science Degree at the School of Engineering and Architecture of the University of Zaragoza since September 2020), and he usually participates regularly in several academic committees.

L'approche européenne de l'encadrement juridique et éthique de l'IA

Olivia Tambou, Maître de Conférences HdR en Droit

Université Paris-Dauphine, PSL
olivia.tambou@dauphine.psl.eu

1. Résumé

En juin 2024, l'Union européenne (UE) a adopté un cadre réglementaire d'envergure pour la mise en circulation de produits et de services d'IA au sein de son marché intérieur. Ce règlement sur l'IA (RIA) a vocation à s'appliquer aussi aux acteurs non européens souhaitant commercialiser leurs produits et services d'IA dans l'UE. Le RIA rend visible à l'échelle mondiale les contours du modèle européen d'encadrement de l'IA. Ce modèle européen repose sur une approche par les risques qui se décline à travers trois types de régimes :

- Premièrement, le RIA fixe une liste exhaustive de huit pratiques interdites d'IA dans l'UE.
- Deuxièmement, le RIA fixe des exigences de conformité préalable à la commercialisation dans l'UE de certains types de produits ou de services d'IA considérés à haut risque.
- Troisièmement, le RIA pose des exigences supplémentaires de transparence pour les modèles d'IA à usage général et pour certains fournisseurs, ou utilisateurs d'IA (IA générative, deep-fake … cf. article 50 RIA)

Si l'objectif du RIA est ambitieux, l'encadrement proposé ne concerne que les usages de l'IA et non directement la recherche théorique en IA. Il s'agit « de promouvoir l'adoption de l'IA axée sur l'humain et digne de confiance tout en garantissant un niveau élevé de protection de la santé, de la sécurité, et des droits fondamentaux consacrés par la charte des droits fondamentaux de l'UE y compris la démocratie, l'état de droit, la protection de l'environnement, de protéger contre les effets néfastes des SIA dans l'Union et de soutenir l'innovation ». (cf. considérant 1 RIA).
L'encadrement européen de l'IA constitue ainsi pour l'UE une nécessité existentielle. Il s'agit d'imposer le respect des valeurs et des droits fondamentaux européens à des acteurs qui ne sont pas toujours européens tout autant que de promouvoir l'émergence d'un écosystème européen intégrant dès la conception ces valeurs et droits fondamentaux. La mise en œuvre du RIA est en conséquence soumise à une gouvernance particulière. Deux nouveautés ont été introduites afin de renforcer la capacité d'expertise et de suivi des autorités chargées de la régulation de l'IA. Premièrement, un bureau de l'IA a été créé au sein de la Commission européenne. Véritable centre d'expertise technologique chargé de surveiller les progrès de l'IA et en particulier des modèles d'IA à usage général, le bureau de l'IA est aussi la cheville ouvrière du contrôle de la mise en œuvre du RIA. Deuxièmement, des bacs à sable réglementaires doivent être créés dans chaque État membre afin de permettre aux autorités nationales compétentes d'accompagner vers la conformité les fournisseurs de systèmes d'IA innovants. Des organismes européens ont été chargés d'adopter des normes techniques européennes pour faciliter la mise en conformité des acteurs. Enfin, la mise en œuvre du RIA comporte une dimension participative. A titre d'exemple un code pratique pour les fournisseurs de modèle d'IA à usage général est en cours d'adoption. Le RIA encourage aussi les fournisseurs de SIA à risque faible à adopter des codes de conduite susceptibles de s'aligner de façon volontaire sur les exigences

des systèmes d'IA à haut risques. Plus globalement, la mise en œuvre du RIA implique l'adoption de règles éthiques par les acteurs pour concrétiser leurs obligations juridiques et définir leurs propres lignes directrices sur l'usage et le développement d'une IA conforme à leurs valeurs.

C'est dire le chantier qui attend l'ensemble des acteurs, jusqu'à l'application pleine et entière du RIA. Actuellement, seules les dispositions relatives aux pratiques interdites les définitions et l'obligation de maîtrise de l'IA (art. 4) sont applicables depuis 2 février 2025. Le régime des modèles d'IA à usage général sera applicable à partir du 2 août 2025, et celui des SIA à haut risque à partir du 2 août 2026.

En attendant, l'encadrement de l'IA dans l'Union européenne repose sur d'autres règles applicables tels que l'encadrement des données liées à l'IA (RGPD, respect droit d'auteur, ouverture et partage de données publiques) et l'encadrement décisions prises par l'IA.

2. Biographie

Olivia Tambou est maître de Conférences HDR en Droit à l'Université Paris Dauphine-PSL où elle enseigne l'éthique et les données aux étudiants de la MIAGE et plus globalement le droit du numérique et de l'IA à des étudiants en droit et en IA au niveau de la Licence et du Master. Elle est aussi membre de l'équipe pédagogique de la PSL week IA responsable. Au sein du Centre de Recherche en Droit Dauphine (CR2D) elle anime un axe de recherche sur le droit de l'IA et l'IA appliquée au Droit. Elle coordonne également une demi-journée dédiée au Droit et à l'éthique au sein de la manifestation annuelle des Dauphine Digital Days. Elle est Fellow de la AI Paris School of AI (PSAI), créée dans le cadre du programme IA Cluster (PR[AI]RIE-PSAI), annoncé dans la stratégie France 2030. Elle a participé en temps d'experte nationale à plusieurs projets européens sur la protection des données à caractère personnel ou le droit de l'IA, tels que Artificial Intelligence and Public Administration, du European Law Institute (2019-2021), PANELFIT, Participatory Approaches to a New Ethical and Legal Framework for ICT (2019), PRODATIES, programme de recherche financé par le ministère de l'économie espagnol sur la mise en œuvre du RGPD. (2019-2021). Elle a été External Scientific Fellow External Scientific Fellow at the MPI Luxembourg for Procedural Law de 2018-2023.

Elle est auteure de quatre ouvrages individuels, 6 directions d'ouvrages collectifs, 22 chapitres d'ouvrages, 22 articles dans des revues internationales avec comités de lecture et de nombreux posts dans des blogs spécialisés. Son dernier ouvrage à paraître chez Bruylant en juin 2025 s'intitule Le Règlement européen sur l'IA et au-delà : quel encadrement de l'IA?

Elle est aussi l'éditrice et la créatrice de blogdroiteuropeen.com.

Vers une sécurité durable des systèmes d'information : le cas de l'optimisation du *vulnerability management* au sein d'un environnement *Cloud*.

Yann Goetgheluck [1,2], **Pierre-Emmanuel Arduin** [2], **Myriam Merad** [1]

1. Université Paris-Dauphine, PSL, LAMSADE UMR CNRS 7243

2. Université Paris-Dauphine, PSL, DRM UMR CNRS 7088

{prenom.nom}@dauphine.psl.eu

RÉSUMÉ. *La sécurité des systèmes d'information (SI) est cruciale pour garantir la continuité des opérations des entreprises. Elle forme l'ossature qui soutient les organisations, semblable à une structure de protection autour de leur colonne vertébrale. C'est une composante du SI qui doit être pérenne pour remplir sa fonction principale tout en prenant en compte les objectifs de responsabilité environnementale des organisations. Un aspect important mais souvent négligé de la sécurité des SI est le processus de gestion des vulnérabilités (Vulnerability Management), qui constitue un lien entre la sécurité du SI et la responsabilité environnementale. En intégrant l'aspect métier des organisations dans la gestion des vulnérabilités, on peut identifier quelles vulnérabilités sont vraiment critiques pour l'organisation concernée. Dans ce travail, nous proposons d'étendre la méthode du Système Commun de Notation des Vulnérabilités (CVSS) afin de mieux prioriser les vulnérabilités et ainsi réduire le nombre de correctifs, de scans et de déploiements qui consomment de l'énergie. Le cas d'une banque, d'un hôpital et d'un gestionnaire de sites web sont abordés afin d'illustrer l'utilisation de la méthode.*

ABSTRACT. *The security of information systems (IS) is crucial to ensuring the continuity of business operations. It is a component of the IS that must be sustainable to fulfill its primary function while considering the organizations environmental responsibility objectives. An important but often overlooked aspect of IS security is the vulnerability management process, which constitutes a link between IS security and environmental responsibility. By integrating the business aspect of organizations into vulnerability management, we can identify which vulnerabilities are truly critical for the concerned organization. In this work, we propose extending the Common Vulnerability Scoring System (CVSS) method to better prioritize vulnerabilities and thus reduce the number of patches, scans, and deployments that consume energy. The cases of a bank, a hospital, and a website manager are discussed to illustrate the use of the method.*

MOTS-CLÉS : *Sécurité des SI, Responsabilité Environnementale, Gestion des Vulnérabilités, CVSS*

KEYWORDS: *IS Security, Environmental Responsibility, Vulnerability Management, CVSS*

1 Introduction

Le système d'information (SI), défini comme un ensemble organisé de ressources (matériel, logiciel, personnel, données, procédures) destiné à gérer l'information au sein des organisations (Reix, 2004), constitue la colonne vertébrale de l'activité immatérielle des entreprises modernes (Legrenzi, 2016). En raison de ce rôle central, il doit être sécurisé, protégé et maintenu opérationnel.

Le système de management de la sécurité de l'information (SMSI) répond à cet impératif en protégeant la confidentialité, l'intégrité et la disponibilité des informations (triade CIA [1]). Les normes ISO/IEC 27001 et 27002 offrent un cadre structuré pour gérer la sécurité de manière continue et adaptée aux risques métiers (**?**Watkins, 2022).

Nous considérons ici le SMSI comme un pilier du SI, bien qu'il n'en couvre pas toutes les dimensions. La gestion des vulnérabilités, par exemple, reste principalement centrée sur les aspects techniques, au détriment des facteurs humains et métiers. Ce travail vise précisément à explorer ces dimensions souvent négligées.

Par ailleurs, si la sécurité et la performance demeurent les priorités dans la conception des systèmes d'information, leur dimension éco-responsable reste encore largement sous-explorée (Chen *et al*., 2008; Mastelic *et al*., 2014). Le concept d'organisation écologiquement durable (Starik, Rands, 1995) appelle pourtant à mobiliser les SI comme leviers de pratiques plus respectueuses de l'environnement. Certaines initiatives, comme celle menée par le Campus Cyber en partenariat avec Wavestone, tentent de mesurer l'empreinte carbone des SMSI (Cyber4Tomorrow, 2025), mais peinent à intégrer d'autres indicateurs environnementaux tout aussi essentiels, tels que la consommation d'eau ou le cycle de vie des ressources numériques. À ce jour, aucune méthodologie éprouvée ne permet de quantifier de manière fiable l'impact environnemental des activités de cybersécurité au sein des organisations. Les travaux de Berthelot *et al*. (2024) montrent que cette lacune existe également à un niveau plus large, concernant l'évaluation de l'impact environnemental d'un ensemble de services numériques. Leur proposition constitue une première méthodologie en ce sens, mais souffre d'une limite majeure : l'absence de référentiels permettant une comparaison significative.

Dans ce contexte, notre recherche s'articule autour de la question suivante : **comment renforcer la sécurité du SI via le *Vulnerability Management* tout en maîtrisant la consommation d'énergie associée au SMSI ?** Nous plaidons pour une amélioration de l'efficience du SMSI, en commençant par sa composante vulnérabilités, et évaluons ses effets sur les ressources à travers trois cas pratiques : une banque, un hôpital et un site vitrine.

1. Nous utiliserons la traduction française : Confidentialité, Intégrité et Disponibilité.

2 Fondements théoriques de la relation complexe mais stratégique entre sécurité et durabilité écologique des SI

La relation entre les systèmes d'information (SI) et l'écologie constitue un domaine de recherche à la fois complexe et prometteur dans le cadre des enjeux de durabilité. Longtemps perçus comme des contributeurs majeurs à l'impact environnemental négatif des organisations, les SI évoluent aujourd'hui pour devenir des leviers stratégiques capables de mesurer, d'analyser et de réduire ces impacts. Comme le suggèrent Chen *et al*. (2008), les SI peuvent agir comme des catalyseurs de durabilité écologique, notamment grâce à des outils avancés de modélisation et de suivi de l'empreinte environnementale.

Cependant, les avantages potentiels des SI dans cette transition écologique sont contrebalancés par leur propre impact énergétique. Wang (2021) propose une approche intégrée via une théorie écologique des écosystèmes d'innovation numérique, où les SI sont intégrés dans une dynamique systémique cherchant à équilibrer la performance technologique et les impératifs écologiques. Pourtant, cet équilibre reste fragile. Par exemple, si la dématérialisation — facilitée par les technologies de télécommunication et le télétravail — réduit la consommation de ressources physiques et les déplacements, la croissance exponentielle des besoins en stockage et en traitement des données entraîne une hausse significative de la consommation énergétique. Ce paradoxe se traduit par l'augmentation de la demande en centres de données sécurisés, indispensables à la continuité des services mais également très consommateurs de ressources matérielles et énergétiques, notamment en raison des exigences croissantes en matière de cybersécurité (Akhter, Othman, 2016).

La gestion de la sécurité de l'information illustre particulièrement bien cette tension entre les bénéfices des SI et les coûts énergétiques. Le renforcement des mécanismes de cybersécurité, bien qu'essentiel, génère souvent une surconsommation d'énergie en raison de l'ajout de dispositifs de protection et de redondances matérielles nécessaires pour garantir la résilience des infrastructures. Par exemple, les centres de données *Cloud*, qui constituent une pierre angulaire de la sécurisation des informations, nécessitent d'importantes ressources énergétiques, malgré les efforts d'optimisation soulignés par Mastelic *et al*. (2014).

Pour atténuer cet impact énergétique, des stratégies prometteuses reposent sur une hiérarchisation plus précise et personnalisée des vulnérabilités, alignée avec les objectifs spécifiques des organisations. Une priorisation rigoureuse permet non seulement une allocation plus efficiente des ressources, mais aussi une réponse adéquate aux exigences croissantes en matière de cybersécurité. À grande échelle, cette approche pourrait conduire à une réduction substantielle de la consommation énergétique des SI. Toutefois, sa mise en œuvre nécessite une réflexion approfondie sur les compromis entre performance écologique et sécurité, ainsi qu'une intégration cohérente des principes de durabilité à long terme dans la gestion des SI.

3 Proposition d'un cadre de recherche exploratoire

Cette étude adopte une approche inductive, l'ITDTA (Inductive Top-Down Theorizing Approach), ancrée dans la tradition pragmatiste (Shepherd, Sutcliffe, 2011). Elle vise à faire émerger des concepts théoriques à partir de l'analyse exploratoire de données empiriques, plutôt qu'à tester des hypothèses préexistantes. Le point de départ est ici le *Vulnerability Management*, utilisé comme prisme pour explorer les impacts de dimensions complémentaires, notamment humaines et métier. Pour classifier les critères de criticité des vulnérabilités, une méthode itérative de type Delphi (Rowe, Wright, 1999) sera mobilisée. Ce processus s'appuie sur les avis d'un panel d'experts, recueillis via plusieurs cycles anonymes afin de converger vers un consensus. L'analyse suivra une logique d'allers-retours entre données et cadres théoriques, assurant une compréhension ancrée dans les réalités observées (Shepherd, Sutcliffe, 2011).

Le Système de Management de la Sécurité de l'Information (SMSI) repose sur trois composantes principales dont les interrelations sont expliquées dans la littérature par (Nyanchama, 2005) et illustrées par la figure (Fig. 1) :

– La gestion des menaces (*Threat Management*) : consiste à identifier et à définir les vecteurs et les types d'attaques susceptibles de cibler les SI.

– La gestion des vulnérabilités (*Vulnerability Management*) : vise à détecter et analyser les failles pouvant être exploitées par ces menaces à des fins malveillantes.

– La gestion des risque (*Risk Management*) : constitue la pierre angulaire du SMSI en utilisant les éléments issues des deux processus précédents pour évaluer les risques pesant sur l'ensemble du SI de l'organisation.

Cependant, dans la littérature scientifique, le SMSI est souvent réduit à sa seule composante de gestion des risques comme en témoigne la figure (Fig. 2) tirée des travaux de Al-Dhahri *et al*. (2017). Cette vision réductrice présente des limites significatives, dont l'une des plus importantes est le manque d'informations fournies par le *Vulnerability Management*. Bien que certaines recherches, telles que celles de (Nyanchama, 2005), tentent de relier le *Vulnerability Management* à l'ensemble des dimensions du SI. En pratique, ce processus est encore largement cantonné à la gestion des systèmes informatiques. Cette focalisation sur le seul aspect technologique conduit à négliger deux dimensions essentielles du SI : l'aspect métier, qui reflète les processus stratégiques des organisations, et l'aspect humain, qui englobe les interactions des individus avec le système. Cette lacune, partagée aussi bien dans les travaux académiques que dans les pratiques des entreprises, limite l'efficacité globale du SMSI et constitue un frein à l'intégration d'une approche plus holistique de la sécurité des systèmes d'information.

Les travaux de Choi, Lee (2015) apportent une contribution novatrice au *Vulnerability Management* en intégrant la dimension métier dans leur approche. Alors que les méthodologies classiques, telles que le *Common Vulnerability Scoring System* (CVSS), se concentrent exclusivement sur les aspects techniques des vulnérabilités (par exemple, le vecteur d'attaque, la complexité, les privilèges nécessaires, etc.), Choi, Lee (2015) introduisent une perspective élargie en prenant en compte les inté-

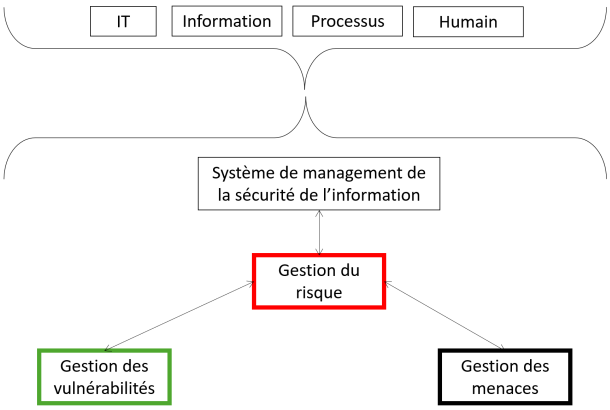


FIGURE 1. *Les composants principaux du SMSI*

rêts stratégiques de l'organisation, définis ici comme la dimension métier du SI. Cette approche permet une hiérarchisation des vulnérabilités mieux adaptée au contexte spécifique du SI, plutôt qu'à un traitement limité au système informatique. En personnalisant l'évaluation des vulnérabilités selon les priorités organisationnelles, il devient possible de mieux aligner les actions de sécurisation avec les besoins réels de l'entreprise.

Pour formaliser cette démarche, (Choi, Lee, 2015) ont proposé un modèle de calcul du score d'importance de l'information basé sur les trois critères fondamentaux de la triade CIA (Confidentialité, Intégrité et Disponibilité). Le score global d'importance est ainsi défini comme la somme des scores obtenus sur chacun des critères :

$$\textit{Information importance score} = \sum \mathrm{C} + \sum \mathrm{I} + \sum \mathrm{A} \tag{1}$$

– **C** représente le score évaluant l'importance de la confidentialité de l'information, c'est-à-dire la capacité à empêcher tout accès ou divulgation non autorisés.

– **I** représente le score mesurant l'importance de l'intégrité de l'information, garantissant qu'elle ne soit ni altérée ni modifiée de manière non autorisée.

– **A** représente le score indiquant l'importance de la disponibilité (*availability*) de l'information, assurant qu'elle reste accessible et utilisable par les personnes autorisées au moment opportun.

Cette méthode introduit de la flexibilité en permettant d'ajuster les pondérations des critères en fonction des spécificités de l'organisation, renforçant ainsi l'efficacité du processus de gestion des vulnérabilités. En intégrant la dimension métier, Choi et Lee démontrent qu'il est possible de dépasser les limites des approches purement

techniques, tout en offrant un cadre adaptable pour répondre aux exigences complexes des systèmes d'information modernes.

Choi, Lee (2015) ont mené un consensus d'experts sur les critères permettant de quantifier l'importance de l'information dans le cadre du *Vulnerability Management*. Ce travail s'appuie sur une intégration des principaux cadres de contrôle de la sécurité, tels que les normes ISO 27 000, le programme CSA STAR *Cloud Security Alliance, Security, Trust, Assurances, and Risk*, les recommandations de l'ENISA (*European Union Agency for Cybersecurity*), ainsi que les directives du BSI (*Bundesamt für Sicherheit in der Informationstechnik*).

– L'***International Organization for Standardization*** 27001 définit les exigences nécessaire à la mise ne place d'un SMSI et fournit un cadre normatif pour garantir la conformité organisationnelle (Julisch, Hall, 2010),

– **CSA STAR** documente les contrôles de sécurité et de confidentialité spécifiques aux environnement du *Cloud computing* en guidant les organisations dans l'évaluation et la gestion des risques liés aux services *Cloud* (Dix, 2012),

– **ENISA** fournit des recommandations et développe des cadres de bonnes pratiques pour renforcer la cybersécurité en Europe, en mettant un accent particulier sur la protection des infrastructures critiques et les standards émergents (Cavelty, Smeets, 2023),

– **BSI** définit les normes et les lignes directrices pour la sécurisation des infrastructures numériques, et la gestion des risques adaptés aux exigences modernes de la cybersécurité (Förderer *et al*., 2019).

Ces référentiels, en combinant des exigences organisationnelles, des bonnes pratiques et des cadres réglementaires, constituent une base solide pour la mise en œuvre et l'amélioration continue d'un SMSI. Ils permettent aux organisations d'identifier, d'évaluer et d'atténuer efficacement les risques liés à la sécurité de l'information. Le travail de Choi, Lee (2015) a permis l'élaboration d'un logiciel opérationnel, testé dans une organisation publique. Ce logiciel, en intégrant les critères définis par les experts et les référentiels mentionnés, propose une hiérarchisation des vulnérabilités plus contextualisée et adaptée aux besoins spécifiques de l'organisation.

Dans cette étude, nous proposons d'étendre cette méthode en la testant sur une infrastructure *Cloud* standard utilisée dans diverses organisations privées. Plus précisément, nous comparerons la hiérarchisation des vulnérabilités qu'elle génère avec les CVSS pour cinq vulnérabilités représentatives des scénarios courants rencontrés en offres SaaS (*Software as a Service*). Ces vulnérabilités, identifiées dans des travaux récents (Abbasi, 2024; Jogi, 2023; Kadu, 2024; Ferguson, 2020), reflètent des enjeux critiques et permettront d'évaluer l'efficacité de la méthode dans des contextes organisationnels variés.

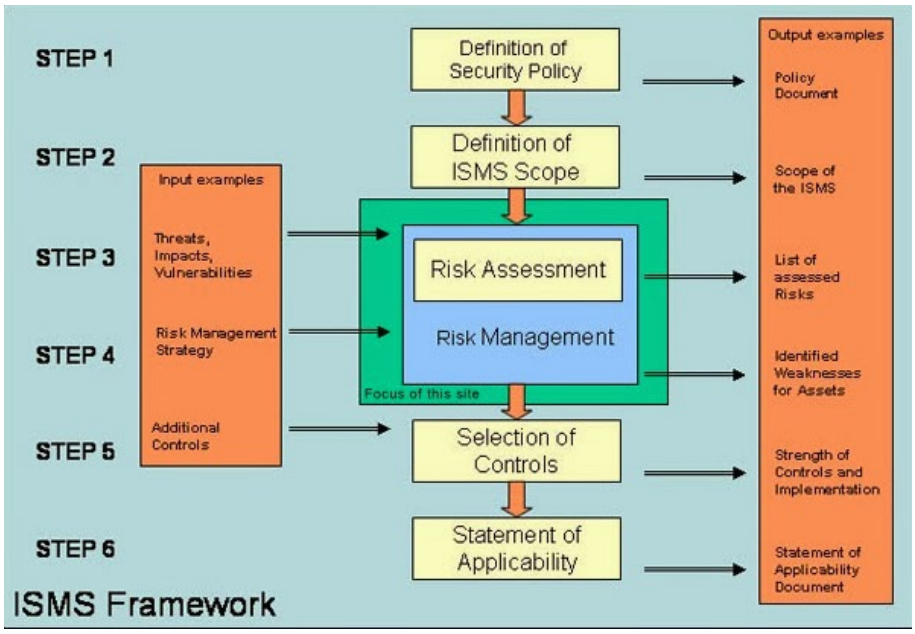


FIGURE 2. *Processus de développement du système de gestion de la sécurité de l'information (Al-Dhahri* et al.*, 2017) (source : http://www.enisa.europa.eu)*

4 Études de cas : 5 vulnérabilités dans une infrastructure *cloud* de 3 organisations différentes

Afin de visualiser et d'évaluer l'intégration de l'aspect métier dans la hiérarchisation des vulnérabilités, nous avons procédé à une comparaison de deux approches . D'une part, une hiérarchisation se fondant exclusivement sur le CVSS, et d'autre part, celle dérivée de la méthode de Choi, Lee (2015). Cette comparaison vise à déterminer s'il existe des différences significatives entre les deux méthodes, et à évaluer leur pertinence respective. Nous avons employé le schéma d'une infrastructure représentative d'un environnement *Cloud* (Fig. 3) dans lequel cinq vulnérabilités informatiques ont été introduites. Cette infrastructure a ensuite été placée dans trois contextes différents : une banque, un hôpital et un site *web* vitrine. Cette approche permet d'étudier deux méthodes d'évaluation des vulnérabilités dans des environnements aux priorités variées. L'objectif est de comparer la méthode de Choi et Lee (2015) à l'approche standard du CVSS, afin d'évaluer dans quelle mesure la prise en compte du contexte métier peut influencer les priorités de gestion des vulnérabilités et améliorer la sécurité des systèmes d'information.

Le schéma illustre le parcours d'un client vers un service *Cloud*. L'accès se fait via Internet ou un VPN, menant à l'espace du fournisseur de services *Cloud* (*Cloud Service Provider*, CSP).On y trouve une infrastructure SaaS (*Software As A Service*) composée de plusieurs éléments : un *Firewall* pour filtrer les accès non autorisés, un

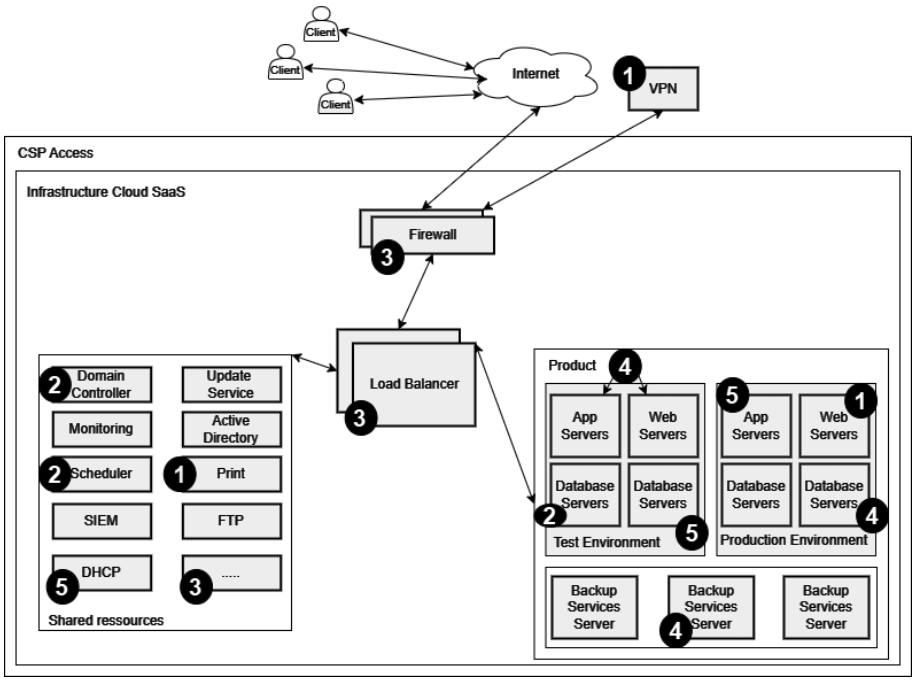


FIGURE 3. *Exemple d'une infrastructure Cloud avec des vulnérabilités détectées*

Load Balancer pour répartir la charge entre les ressources mutualisées, et le produit final qui intègre l'application, un serveur de base de données pour stocker les informations, et un serveur de sauvegarde en cas de panne ou de cyberattaque. Enfin, la zone des ressources partagées regroupe les outils nécessaires à la gestion et à l'optimisation du service.

Pour accéder aux services proposés, l'utilisateur doit franchir plusieurs niveaux de protection et de gestion des accès. La première ligne de défense est le *firewall*, élément fondamental de la sécurité réseau, qui filtre les connexions aux différentes adresses IP et empêche tout accès non autorisé. Une fois cette vérification effectuée, le *LoadBalancer* intervient pour rediriger l'utilisateur vers la ressource cible, qu'il s'agisse d'un serveur *web* ou d'une application spécifique.

Cet écosystème repose sur une orchestration de multiples ressources ayant chacune un rôle précis dans la gestion et l'exploitation des services. Les administrateurs s'appuient sur ces ressources afin de garantir la continuité, la disponibilité et la sécurité des services, conformément aux exigences organisationnelles (Nyanchama, 2005; Akhter, Othman, 2016; Choi, Lee, 2015; Albaroodi, Anbar, 2024).

Dans ce contexte et à partir de cet exemple (Fig. 3), nous avons intégré plusieurs vulnérabilités informatiques. Ces vulnérabilités permettent d'analyser leur positionnement et leur interaction avec les différentes composantes de l'infrastructure *Cloud*,

tout en testant différentes méthodologies de hiérarchisation de la criticité, notamment le CVSS et la méthode de Choi, Lee (2015). Cette approche met en évidence l'impact des vulnérabilités sur les infrastructures modernes et souligne la nécessité d'une gestion des risques adaptée.

4.1 Des vulnérabilités représentatives

Pour cette étude, cinq vulnérabilités ont été sélectionnées pour représenter un éventail large de menaces pouvant affecter l'infrastructure *Cloud*, indépendamment de l'organisation concernée (Rotlevi, 2025; Smith, 2023; owasp.org, 2024).

1. `CVE-2021-22965` (Spring4Shell) - **CVSS 7.5** (*High*) : Permet l'exécution de code à distance via une mauvaise gestion des requêtes HTTP dans le *framework* Spring.

2. `CVE-2021-44228` (Log4Shell) - **CVSS 10** (*Critical*) : Permet l'exécution de code arbitraire via l'exploitation malveillante de la journalisation dans Log4j.

3. `CVE-2020-10188` - **CVSS 9.8** (*Critical*) : Permet l'exécution de commandes malveillantes sur des systèmes F5 BIG-IP vulnérables.

4. `CVE-2020-11023` - **CVSS 6.1** (*Medium*) : Permet l'injection de scripts malveillants (XSS) dans des pages *web*.

5. `CVE-2020-2551` - **CVSS 9.8** (*Critical*) : Permet l'exécution de requêtes SQL malveillantes pour accéder ou manipuler des données sensibles.

En se basant uniquement sur le score CVSS, qui reflète une hiérarchisation standard sans prise en compte du contexte organisationnel, les vulnérabilités se classent, quelle que soit l'organisation, comme suit *Forum of Incident Response and Security Teams* (FIRST, 2023) :

1. **CVSS 10** (*Critical*) : (2) `CVE-2021-44228`
2. **CVSS 9.8** (*Critical*) : (3) `CVE-2020-10188`
3. **CVSS 9.8** (*Critical*) : (5) `CVE-2020-2551`
4. **CVSS 7.5** (*High*) : (1) `CVE-2021-22965`
5. **CVSS 6.1** (*Medium*) : (4) `CVE-2020-11023`

Ainsi, selon cette hiérarchisation, la vulnérabilité (2) – Log4Shell – serait considérée comme la plus prioritaire, car elle est jugée critique, tandis que la vulnérabilité (4) – une faille XSS – serait reléguée au dernier rang, en raison de son score relativement faible. Cette classification repose uniquement sur des critères techniques définis par le CVSS, sans prendre en compte le contexte d'exploitation ou l'impact métier.

Cependant, en intégrant l'aspect métier d'une organisation, comme une banque, un hôpital ou un gestionnaire de sites *web* vitrine, l'ordre de priorité peut être radicalement modifié. Par exemple, une vulnérabilité impactant fortement la disponibilité pourrait être prioritaire pour un hôpital, tandis qu'une faille compromettant la confidentialité serait plus préoccupante pour une banque.

Afin d'évaluer ces différences d'impact, nous avons intégré ces cinq vulnérabilités dans l'infrastructure *Cloud* étudiée précédemment, en conservant les numéros attribués à chaque vulnérabilité avant leur hiérarchisation (voir Fig. 3). Cette approche permet de démontrer l'importance d'adapter les stratégies de remédiation aux contextes organisationnels spécifiques et de ne pas se baser uniquement sur les scores CVSS.

4.2 Scores de vulnérabilité et discussions

D'après les travaux de Choi, Lee (2015) et en appliquant la formule présentée dans la section 3, adaptée aux spécificités des organisations et aux priorités variables attribuées à l'importance de l'information dans chaque secteur (voir Annexe. 5), nous avons établi le tableau récapitulatif des résultats (Tab. 1) :

TABLEAU 1. *Notation de criticité des vulnérabilités par secteur en utilisant la méthode de Choi, Lee (2015).*

Vulnérabilité	**Organisation**	**∑C**	**∑I**	**∑A**	**Total**
CVE-2021-22965	Banque	13	11	8	32
	Hôpital	10	13	13	36
	Site *web* Vitrine	6	6	12	24
CVE-2021-44228	Banque	16	13	10	39
	Hôpital	10	14	15	39
	Site *web* Vitrine	8	9	6	23
CVE-2020-10188	Banque	13	11	9	33
	Hôpital	11	12	14	37
	Site *web* Vitrine	7	8	6	21
CVE-2020-11023	Banque	8	9	7	24
	Hôpital	7	11	12	30
	Site *web* Vitrine	5	6	4	15
CVE-2020-2551	Banque	15	14	10	39
	Hôpital	9	14	14	37
	Site *web* Vitrine	7	8	6	21

Pour obtenir les chiffres présentés dans cette étude, nous nous sommes appuyés sur une analyse approfondie d'articles représentant les points de vue de différents secteurs : les banques (Dumalanede, 2019; Lobez, Vilanova, 2006; Bobillier-Chaumon *et al*., 2006), les hôpitaux (Frenkiel *et al*., 2007; Juven, 2013) et les sites *web* vitrine (Stephane, 2020; Dirigeant, 2024). L'analyse a été structurée autour de la triade de sécurité de l'information – *Confidentialité, Intégrité et Disponibilité* – à travers des critères spécifiques.

Concernant la **confidentialité**, nous avons évalué la sensibilité des informations, la présence de restrictions d'accès et la nécessité de leur protection. Pour l'**intégrité**, les critères incluaient la capacité de restreindre les modifications, la fréquence des

sauvegardes et l'importance des audits des changements. Enfin, la **disponibilité** a été mesurée en fonction de la nécessité d'un accès continu et de l'impact des interruptions potentielles sur l'organisation.

Ces critères ont permis d'évaluer et de hiérarchiser les vulnérabilités identifiées, bien que les résultats obtenus soient limités par l'absence d'un consensus d'experts, ce qui constitue une des limites notables de cette étude. Les scores obtenus reflètent néanmoins des tendances significatives. Pour la vulnérabilité (1), qui concerne principalement les *VPN*, les scores étaient de 32 sur 40 pour une banque, 36 pour un hôpital et 24 pour un site *web* vitrine. Cette vulnérabilité, qui pourrait permettre une exécution de code à distance compromettant l'ensemble de l'infrastructure, est particulièrement critique pour un hôpital en raison des exigences accrues en matière de disponibilité et d'intégrité des données.

Pour la vulnérabilité (2), touchant les *Domain Controllers*, la banque et l'hôpital obtiennent un score de 39 sur 40, tandis que le site *web* vitrine atteint 23 sur 40. Cette disparité s'explique par l'importance légale et économique des informations manipulées par les banques et les hôpitaux. Pour la vulnérabilité (3), liée au *LoadBalancer*, les scores sont respectivement de 33 pour la banque, 37 pour l'hôpital et 21 pour le site *web* vitrine. La disponibilité étant prioritaire pour un hôpital, cette vulnérabilité y est plus critique que pour une banque, où la protection des données est prioritaire.

La vulnérabilité (4), avec des scores de 24 pour la banque, 30 pour l'hôpital et 15 pour le site *web* vitrine, présente un impact limité en raison de son périmètre restreint. Enfin, pour la vulnérabilité (5), les scores sont de 39 pour la banque, 37 pour l'hôpital et 21 pour le site *web* vitrine, reflétant l'importance stratégique de ces vulnérabilités malgré leur faible exploitabilité.

L'analyse révèle des hiérarchisations différentes selon la méthode utilisée. Avec la méthode de Choi, Lee (2015), certaines vulnérabilités critiques selon le CVSS, comme la vulnérabilité (1), sont reléguées en seconde position pour la banque et le site *web* vitrine, tout en restant prioritaires pour l'hôpital. Les vulnérabilités (3) et (5) présentent également des différences de classement, mais leurs scores restent globalement cohérents entre les deux approches. Cela souligne que les deux méthodes, bien qu'indépendantes, sont complémentaires. Le CVSS, orienté vers l'aspect technique, et la méthode de Choi, Lee (2015), centrée sur les priorités métier, apportent des perspectives distinctes qui enrichissent la compréhension et la gestion des vulnérabilités.

Toutefois, ces approches présentent des limites. En examinant les interactions entre vulnérabilités, il apparaît que certaines d'entre elles, comme la vulnérabilité (1), sont critiques uniquement dans des contextes spécifiques (par exemple, le *VPN*). De même, la vulnérabilité (2) n'impacte une organisation qu'à travers les *Domain Controllers*, et la vulnérabilité (3) devient critique parce qu'elle affecte le *LoadBalancer*. Les vulnérabilités (4) et (5) ont un impact limité en raison de leur faible exploitabilité. Cette contextualisation montre que la hiérarchisation doit prendre en compte les interdépendances et le chaînage des vulnérabilités. Si cela est gérable dans des infrastructures de taille modeste, cela devient rapidement impraticable pour des organisations com-

plexes avec plusieurs milliers de serveurs et des infrastructures variées. L'utilisation d'une méthode industrialisée devient alors obligatoire.

Ainsi, bien que la vulnérabilité (3) représente un danger critique en raison de son impact sur des éléments essentiels de l'infrastructure *Cloud*, elle n'arrive qu'en seconde position avec le CVSS et en quatrième position selon la méthode de Choi, Lee (2015). Ce cas illustre l'influence de l'impact métier sur la priorisation des vulnérabilités et met en lumière les limites des méthodes existantes, qui ne considèrent qu'un aspect du *SI* : le CVSS se concentre sur les aspects techniques, tandis que la méthode de Choi, Lee (2015) privilégie l'importance métier et la valeur de l'information. Ces observations renforcent l'idée qu'un *SMSI* intégré et adapté, prenant en compte la globalité du *SI*, est nécessaire pour une gestion durable et efficace des vulnérabilités.

Trois approches distinctes apparaissent : la méthode standard basée sur les CVSS, une méthode personnalisée vue chez Chois et Lee (2015), et une autre qui exploite le chaînage des vulnérabilités, toutes montrant des différences de priorisation dans ce cas pratique, comme le montre la Fig. 4. Des lacunes persistent donc dans le développement d'une méthode complète et adaptable. Il est nécessaire d'identifier ces lacunes avec précision et de trouver un moyen de les combler.

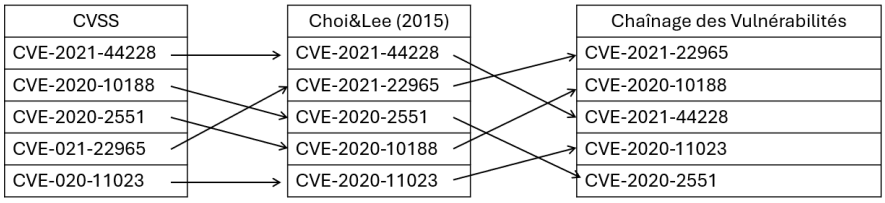


FIGURE 4. *Différence de hiérarchisation de vulnérabilité d'après 3 méthodes différentes*

En prenant en compte le chaînage des vulnérabilités, certaines failles critiques peuvent devenir totalement inexploitables, ne posant alors aucun danger pour les activités de l'organisation. Il ne s'agit donc pas d'ignorer des vulnérabilités sous prétexte d'économiser de l'énergie, mais plutôt de déterminer précisément si une vulnérabilité représente réellement un risque pour l'organisation et d'agir en conséquence.

Limites, conclusions et perspectives

Les méthodes existantes, telles que le *CVSS*, qui s'impose comme un standard de référence pour évaluer la criticité des vulnérabilités, et la méthode de Choi, Lee (2015), qui intègre les dimensions métiers, constituent des bases solides pour la gestion des vulnérabilités. Toutefois, ces approches présentent des limites importantes lorsqu'elles sont appliquées à de grandes organisations aux structures complexes et diversifiées. Le *CVSS* se concentre exclusivement sur les aspects techniques des vulnérabilités, négligeant les spécificités métiers des organisations. Inversement, la méthode de Choi, Lee (2015), bien qu'innovante sur le plan métier, omet souvent les contraintes techniques qui restent essentielles dans les environnements hautement technologiques.

Ce travail de recherche qui se poursuit vise à combiner ces trois perspectives en développant une méthode personnalisée de *Vulnerability Management*. Capable d'intégrer de manière équilibrée les dimensions métier et technique propres à chaque organisation tout en prenant en compte le chaînage des vulnérabilités. L'objectif est de fournir une approche optimisée de la hiérarchisation des vulnérabilités, réduisant ainsi les besoins en remédiation et contribuant à une forme de sobriété numérique dans le domaine de la sécurité des *SI*.

Dans cette optique, nous prévoyons de mobiliser la méthode *Delphi* pour affiner notre démarche. En impliquant un panel d'experts issus de divers secteurs, nous chercherons à établir un véritable consensus sur les critères essentiels à considérer, notamment l'importance de l'information et les priorités spécifiques des parties prenantes. Cette approche permettra de concevoir un cadre plus robuste et adaptable pour répondre aux besoins variés des entreprises. Un travail pour trouver ou déterminer une métrique propre à l'empreinte écologique de la sécurité des systèmes d'information est prévu afin de pouvoir évaluer la pertinence de ces travaux. L'ambition à long terme est de proposer une solution personnalisable, complète, durable et efficiente pour la gestion des vulnérabilités en entreprise. Cette méthode cherche également à transformer la perception du SMSI, souvent considérée comme un centre de coûts, en un véritable investissement stratégique. Une gestion efficiente des vulnérabilités peut en effet améliorer non seulement la résilience des organisations face aux menaces, mais également leur image de marque et leur engagement en faveur d'une sobriété numérique durable. En intégrant des considérations économiques et environnementales, cette approche vise à concilier efficacité opérationnelle, viabilité économique et durabilité écologique.

Bibliographie

Abbasi S. (2024, 11). *Qualys TRU Uncovers Five Local Privilege Escalation Vulnerabilities in needrestart | Qualys Security Blog.* Consulté sur https://blog.qualys.com/vulnerabilities-threat-research/2024/11/19/qualys-tru-uncovers-five-local-privilege-escalation-vulnerabilities-in-needrestart

Akhter N., Othman M. (2016). Energy aware resource allocation of cloud data center: review and open issues. *Cluster Computing*, vol. 19, n° 3, p. 1163–1182.

Albaroodi H., Anbar M. (2024). *Journal of Applied Data Sciences*, vol. 6, n° 1, p. 155–177. Consulté sur https://bright-journal.org/Journal/index.php/JADS/article/view/324

Al-Dhahri S., Al-Sarti M., Abdaziz A. (2017). Information security management system. *International Journal of Computer Applications*, vol. 158, p. 29-33.

Berthelot A., Caron E., Laage R. de, Lefèvre L., Nicolas A. (2024). *Fine-grained methodology to assess environmental impact of a set of digital services.* Consulté sur https://hal.science/hal-04928998 (working paper or preprint)

Bobillier-Chaumon M.-E., Dubois M., Retour D. (2006). *L'acceptation des nouvelles technologies d'information : le cas des systèmes d'information en milieu bancaire.* Consulté sur

https://shs.hal.science/halshs-01562077v1

Cavelty M. D., Smeets M. (2023). Regulatory cybersecurity governance in the making: the formation of enisa and its struggle for epistemic authority. *Journal of European Public Policy*, vol. 30, nº 7, p. 1330–1352. Consulté sur https://doi.org/10.1080/13501763.2023.2173274

Chen A. J., Boudreau M.-C., Watson R. T. (2008). Information systems and ecological sustainability. *Journal of Systems and Information Technology*, vol. 10, nº 3, p. 186–201.

Choi M., Lee C. (2015). Information security management as a bridge in cloud systems from private to public organizations. *Sustainability*, vol. 7, nº 9, p. 12032–12051.

Cyber4Tomorrow. (2025, 4). *Présenter la méthodologie d'évaluation empreinte carbone de la cybersécurité - Cyber4Tomorrow*. Consulté sur https://cyber4tomorrow.fr/actions/evaluation-empreinte-carbone-de-la-cybersecurite/

Dirigeant L. B. du. (2024, 11). *Le site vitrine : Définition et utilité pour votre entreprise en 2025*. Consulté sur https://www.leblogdudirigeant.com/site-vitrine-entreprise

Dix J. (2012, Mar 26). Push your cloud supplier to participate in csa star. *Network World*, vol. 29, nº 6, p. 5. Consulté sur https://www.proquest.com/trade-journals/push-your-cloud-supplier-participate-csa-star/docview/1009899414/se-2 (Copyright - Copyright Network World Inc. Mar 26, 2012; Last updated - 2017-11-19)

Dumalanede C. (2019, 12). *Un management stratégique dédié à la prestation de services de santé primaires aux plus démunis des régions en développement : un business model Bottom the Pyramid (BoP) et son système propositionnel*. Consulté sur https://theses.hal.science/tel-03419273v1

Ferguson D. (2020, 10). *Qualys WAS Engine 8.3 released | Qualys Notifications*. Consulté sur https://notifications.qualys.com/product/2020/09/04/qualys-was-engine-8-3-released

FIRST. (2023). *Forum of incident response and security teams, common vulnerability scoring system (cvss) version 3.0*. Consulté sur https://www.first.org/cvss/v4.0/specification-document (Accessed: 2025-02-14, Also available in NVD: https://nvd.nist.gov/)

Frenkiel J., BOUAM S., Triadou P. (2007, 06). L'information en milieu hospitalier : apports potentiels de la qualitique et de la productique. l'exemple de la méthode primaq (production de l'information médicale en assurance qualité). *Santé et systémique*, vol. 10.

Förderer K., Lösch M., Növer R., Ronczka M., Schmeck H. (2019). Smart meter gateways: Options for a bsi-compliant integration of energy management systems. *Applied Sciences*, vol. 9, nº 8. Consulté sur https://www.proquest.com/scholarly-journals/smart-meter-gateways-options-bsi-compliant/docview/2331407740/se-2

Jogi B. (2023, 1). *CVE-2021-244228: Apache Log4j2 Zero Day Exploited in the Wild (Log4Shell) | Qualys Security Blog*. Consulté sur https://blog.qualys.com/vulnerabilities-threat-research/2021/12/10/apache-log4j2-zero-day-exploited-in-the-wild-log4shell

Julisch K., Hall M. (2010, 11). Security and Control in the Cloud. *Information Security Journal A Global Perspective*, vol. 19, nº 6, p. 299–309. Consulté sur https://doi.org/10.1080/19393555.2010.514654

Juven P.-A. (2013, 1). Produire l'information hospitalière. *Revue d anthropologie des connaissances*, vol. 7, nº 4. Consulté sur https://doi.org/10.3917/rac.021.0815

Kadu H. (2024, 3). *March 2024 Web application vulnerabilities released | Qualys notifications*. Consulté sur https://notifications.qualys.com/product/2024/03/29/march-2024-web-application-vulnerabilities-released

Legrenzi C. (2016). Informatique, numérique et système d'information: définitions, périmètres, enjeux économiques. *Vie & Sciences de L'Entreprise*, nº 200, p. 49–76.

Lobez F., Vilanova L. (2006). *La banque productrice d'information*. Paris, France, Presses Universitaires de France eBooks.

Mastelic T., Oleksiak A., Claussen H. *et al*. (2014). Cloud computing: Understanding infrastructure energy consumption for cloud environments. *Future Generation Computer Systems*, vol. 37, p. 101–112.

Nyanchama M. (2005). Enterprise vulnerability management and its role in information security management. *Information Systems Security*, vol. 14, p. 29–56.

owasp.org. (2024). *Owasp top ten | owasp foundation.* Consulté sur https://owasp.org/www-project-top-ten/

Reix R. (2004). *Systèmes d'information et management des organisations* (5e éd.). Paris, France, Vuibert.

Rotlevi S. (2025, 1). *The Basics of AWS Infrastructure Security.* Consulté sur https://www.wiz.io/blog/aws-infrastructure-security-basics

Rowe G., Wright G. (1999). The delphi technique as a forecasting tool: issues and analysis. *International Journal of Forecasting*, vol. 15, no 4, p. 353–375.

Shepherd D. A., Sutcliffe K. M. (2011). Inductive top-down theorizing: A source of new theories of organization. *Academy of Management Review*, vol. 36, no 2, p. 361–380.

Smith T. (2023, 4). *Cybersecurity Risk Fact : Infrastructure Misconfigurations Open the Door to Ransomware | Qualys Security Blog.* Consulté sur https://blog.qualys.com/vulnerabilities-threat-research/2023/04/03/risk-fact-5-infrastructure-misconfigurations-open-the-door-to-ransomware

Starik M., Rands G. P. (1995). Weaving an integrated web: Multilevel and multisystem perspectives of ecologically sustainable organizations. *Academy Of Management Review*, vol. 20, no 4, p. 908–935.

Stephane. (2020, 7). *Pourquoi avoir un site vitrine pour votre entreprise ?* Consulté sur https://management-digital.com/blog/formation/pourquoi-avoir-un-site-vitrine-pour-votre-entreprise/

Wang P. (2021). Connecting the parts with the whole: Toward an information ecology theory of digital innovation ecosystems. *MIS Quarterly*, vol. 45, no 1, p. 397–422.

Watkins S. G. (2022). *ISO/IEC 27001:2022*. Ely, Cambridgeshire, Royaume-Uni, IT Governance Publishing Ltd. Consulté sur https://doi.org/10.2307/j.ctv30qq13d

5 Annexe

TABLEAU 2. *Matrice des critères de sécurité.*

Critère	**Confidentialité (C)**	**Intégrité (I)**	**Disponibilité (D)**
Sensibilité des données impactées	✓	-	-
Nécessité de protéger l'information	✓	-	-
Possibilité de divulgation non autorisée	✓	-	-
Niveau de confiance dans les données	-	✓	-
Risque de modification non autorisée	-	✓	-
Nécessité de validation de l'intégrité	-	✓	-
Besoin d'accès continu aux informations	-	-	✓
Impact en cas d'indisponibilité	-	-	✓
Priorité de récupération après incident	-	-	✓

Chaque vulnérabilité est ensuite analysée en fonction de ces critères.

Calculs détaillés des scores exemple avec : CVE-2021-22965 (Spring4Shell)

Banque

Confidentialité (C)	**Intégrité (I)**	**Disponibilité (D)**
Sensibilité des données impactées: 4	Niveau de confiance dans les données: 4	Besoin d'accès continu aux informations: 3
Nécessité de protéger l'information: 5	Risque de modification non autorisée: 4	Impact en cas d'indisponibilité: 3
Possibilité de divulgation non autorisée: 4	Nécessité de validation de l'intégrité: 3	Priorité de récupération après incident: 2
Score total: $C = 4 + 5 + 4 = 13$	**Score total:** $I = 4 + 4 + 3 = 11$	**Score total:** $D = 3 + 3 + 2 = 8$

Total final: $\sum C + \sum I + \sum D = 32$

Hôpital

Confidentialité (C)	**Intégrité (I)**	**Disponibilité (D)**
Sensibilité des données impactées: 3	Niveau de confiance dans les données: 5	Besoin d'accès continu aux informations: 5
Nécessité de protéger l'information: 4	Risque de modification non autorisée: 4	Impact en cas d'indisponibilité: 4
Possibilité de divulgation non autorisée: 3	Nécessité de validation de l'intégrité: 4	Priorité de récupération après incident: 4
Score total: $C = 3 + 4 + 3 = 10$	**Score total:** $I = 5 + 4 + 4 = 13$	**Score total:** $D = 5 + 4 + 4 = 13$

Total final: $\sum C + \sum I + \sum D = 36$

Site Web Vitrine

Confidentialité (C)	**Intégrité (I)**	**Disponibilité (D)**
Sensibilité des données impactées: 2	Niveau de confiance dans les données: 2	Besoin d'accès continu aux informations: 5
Nécessité de protéger l'information: 2	Risque de modification non autorisée: 2	Impact en cas d'indisponibilité: 4
Possibilité de divulgation non autorisée: 2	Nécessité de validation de l'intégrité: 2	Priorité de récupération après incident: 3
Score total: $C = 2 + 2 + 2 = 6$	**Score total:** $I = 2 + 2 + 2 = 6$	**Score total:** $D = 5 + 4 + 3 = 12$

Total final: $\sum C + \sum I + \sum D = 24$

Systèmes d'Information Responsables : Regards et Perspectives de l'Intérieur

Chloe Gobillot, Rébecca Deneckère

Centre de Recherche en Informatique
Université Paris 1 Panthéon-Sorbonne
rebecca.deneckere@univ-paris1.fr

REFERENCE DE L'ARTICLE INTERNATIONAL Cet article est une synthèse de l'article :
Chloe Gobillot, Rébecca Deneckère: Sustainable Information Systems: Insights from Inside. KES 2024: 2194-2204

1. Introduction

Cette étude se situe à l'intersection des technologies numériques et de la durabilité environnementale, en se concentrant sur le concept de responsabilité numérique. À travers des entretiens avec diverses organisations, nous avons cherché à identifier les perspectives, défis et pratiques liés à la responsabilité numérique. La recherche révèle que, bien que la numérisation soit souvent perçue comme une solution de croissance, elle pose également des défis environnementaux significatifs, contribuant aux émissions de gaz à effet de serre et à la consommation de ressources. Les organisations reconnaissent l'importance de ces enjeux et ont initié diverses pratiques éco-responsables telles que la numérisation responsable et la sobriété numérique. Cependant, la mise en œuvre efficace de ces pratiques nécessite des stratégies contextualisées adaptées aux besoins et cultures organisationnels. Cette étude souligne la nécessité d'une amélioration continue et d'un suivi des pratiques éco-responsables pour atténuer l'impact environnemental des technologies numériques.

2. Éléments clefs des Systèmes d'Information Responsables

Nous avons mené des entretiens avec diverses organisations, y compris des entreprises, des ONG et des associations, pour recueillir des perspectives sur la responsabilité numérique. Les entretiens ont révélé que la numérisation responsable et la sobriété numérique sont des stratégies clés pour réduire l'empreinte environnementale des activités numériques. Cependant, la compréhension actuelle de la responsabilité numérique est encore en évolution, avec des termes variés tels que Green IT ou Green IS qui ajoutent à la complexité. Les organisations sont

conscientes de la nécessité d'évoluer vers des pratiques plus durables, mais elles peuvent manquer de stratégies de mise en œuvre claires. La recherche présentée dans cet article a suivi une méthodologie structurée composée d'entretiens qui a permet d'identifier les éléments principaux de notre analyse.

Les entretiens ont été menés en 2023 avec 8 entités ayant initié des initiatives éco-responsables ou offrant des services connexes, permettant une compréhension approfondie des défis et des niveaux d'engagement au sein des organisations. Ces entités comprenaient des organisations publiques (la Cour des comptes, la région Nouvelle Aquitaine et la CAF de Seine Maritime), une association (Fresque du numérique) et des entreprises privées (Key4Events, Infogreen Factory, Renault et une autre entreprise qui n'a pas souhaité être mentionnée). Tous les entretiens ont suivis un guide d'interview, ont été enregistrés et retranscris.

Voici la liste des éléments clefs identifiés suite à l'analyse des entretiens : Coordination et Stratégie Organisationnelle, Indicateurs de Performance et Normes, Allocation Budgétaire et Planification des Ressources, Développement des Compétences et Sensibilisation, Partenariats et Collaborations, Infrastructure et Gestion des Données, Pratiques de Réemploi et de Recyclage. Ces éléments sont détaillés dans l'article original et ont également été validés par les organisations.

3. Conclusion

Cette étude montre l'importance de la responsabilité numérique comme un enjeu collectif nécessitant des efforts concertés pour intégrer les pratiques éco-responsables dans les normes organisationnelles. Les organisations interrogées soulignent la nécessité d'efforts soutenus et à long terme, y compris des études de recherche et développement pour mieux répondre aux besoins et attentes organisationnels. La responsabilité numérique ne se limite pas aux départements informatiques, mais doit être intégrée dans les politiques de responsabilité sociale des entreprises et impliquer plusieurs départements.

Bien que la responsabilité numérique offre des opportunités significatives, son adoption généralisée fait face à des obstacles en termes de sensibilisation, de priorisation et d'intégration dans les objectifs organisationnels. Les mesures gouvernementales et réglementaires peuvent jouer un rôle crucial, tout comme le soutien et les outils fournis aux organisations. Cette recherche suggère des perspectives supplémentaires sur la gestion du changement et des initiatives collaboratives entre organisations pour promouvoir l'échange de connaissances et une mise en œuvre efficace.

Développement Logiciel Éco-Responsable : Guide pour des Pratiques Durables

Ryan Vernex, Rébecca Deneckère

Centre de Recherche en Informatique
Université Paris 1 Panthéon-Sorbonne, Paris, France
Rebecca.Deneckere@univ-paris1.fr

REFERENCE DE L'ARTICLE INTERNATIONAL Cet article est une synthèse de l'article : Ryan Vernex, Rébecca Deneckère: Eco-Conscious Software Development: A Comprehensive Guide for Sustainable Practices. DB&IS 2024: 18-33

1. Introduction

Selon l'Organisation mondiale de la santé, le changement climatique constitue une menace significative pour la santé mondiale et la stabilité environnementale. L'émission incontrôlée de gaz à effet de serre, principalement due aux activités humaines, accélère le réchauffement climatique et entraîne des perturbations climatiques graves. Ces changements n'affectent pas seulement les écosystèmes naturels, mais impactent également les sociétés humaines, rendant impératif l'adoption de pratiques durables dans tous les secteurs. Cette recherche se concentre sur l'orientation du processus de développement des applications vers une approche durable. En intégrant des principes d'éco-conception et en utilisant des méthodologies avancées, il est possible de créer des logiciels et du matériel qui ne sont pas seulement efficaces, mais aussi respectueux de l'environnement. À travers une analyse approfondie de la littérature existante et d'entretiens avec quatre experts de divers horizons (un manager « Green IT », un développeur Full-Stack, un analyste « Sustainable IT », un consultant), ce travail vise à fournir un guide pour les développeurs et les organisations afin d'adopter des pratiques informatiques durables, contribuant ainsi à un avenir plus vert et plus résilient.

2. Une Approche Multi-Niveaux pour Améliorer la Responsabilité

Pour répondre à la question de recherche *Comment orienter le processus de développement des applications vers une approche durable ?* nous proposons une approche multi-niveaux, comme illustré dans la figure 1. Cette approche met en lumière les interactions entre différents niveaux : le niveau du modèle d'affaires, le niveau de la gestion, et le niveau du développement. En partant du haut vers le bas,

nous identifions d'abord les opportunités d'amélioration au niveau du modèle d'affaires, en analysant de manière critique les modèles existants pour les aligner avec les objectifs de durabilité. Une fois les modèles appropriés choisis, la transition vers ces nouveaux modèles est initiée, marquant le début de la mise en œuvre.

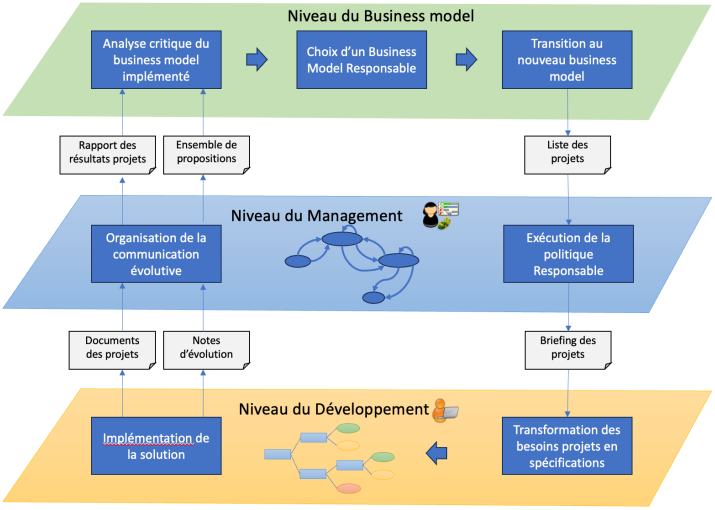


Figure 1. Approche multi-niveaux

Au niveau de la gestion, les équipes organisent les projets initiaux et allouent les ressources nécessaires. Ensuite, la tâche est transmise aux équipes de développement, qui intègrent les pratiques durables dans les spécifications des projets. Ce processus itératif permet une amélioration continue, en s'assurant que les idées et les retours d'expérience des équipes de développement sont pris en compte pour affiner les modèles d'affaires. L'équipe de mise en œuvre joue un rôle crucial en facilitant la communication entre les différents niveaux, garantissant ainsi que les perspectives des équipes techniques soient intégrées dans les décisions stratégiques. Ce processus forme une boucle continue d'amélioration, où chaque niveau contribue à l'optimisation des pratiques durables dans le développement des applications.

3. Conclusion

Nous avons validé le guide par une série d'entretiens supplémentaires pour recueillir les perspectives des experts sur les directives proposées. Tous les interviewés ont approuvé la proposition de figure multi-niveaux, soulignant l'importance d'une vue d'ensemble du problème et la pertinence du cycle proposé. Ce guide met en avant l'importance d'utiliser un modèle d'affaires durable pour guider une approche durable. Il examine de plus près le processus de gestion des exigences d'impact au niveau de la gestion et offre des lignes directrices pour la mise en œuvre de solutions, qu'il s'agisse de transformer une solution existante ou d'en implémenter une nouvelle.

Apprentissage par renforcement pour la personnalisation de l'UX dans les Écosystèmes d'Affaires Numériques

Mustapha Kamal BENRAMDANE [1], **Elena KORNYSHOVA** [1]

CEDRIC, CNAM, 292 rue Saint Martin, 75003 Paris, France
mustapha-kamal.benramdane@lecnam.net,elena.kornyshova@cnam.fr

RÉFÉRENCE DE L'ARTICLE INTERNATIONAL. Cet article est un résumé de l'article: Benramdane M. K., Kornyshova E. (2024). Reinforcement Learning to personalize User eXperience within Digital Business Ecosystems, COMPSAC 2024, p. 584–593, Osaka Japan.

1. Introduction

L'article propose une représentation des données et leur utilisation par un système de recommandation basé sur des algorithmes d'apprentissage par renforcement (RL - Reinforcement Learning) au sein des Ecosystèmes d'Affaires Numériques (DBE - Digital Business ecosystem) (Benramdane, Kornyshova, 2024). Nous nous intéressons à l'impact de la recommandation sur la satisfaction des utilisateurs au sein de ces écosystèmes. Le principe fondamental qui sous-tend cette recherche est la reconnaissance que l'expérience utilisateur, les intentions des utilisateurs et les données contextuelles au sein des plateformes numériques sont cruciales pour une meilleure recommandation. Les résultats de l'expérience utilisateur (UX - User eXperience), la satisfaction ou la frustration de l'utilisateur sont des paramètres qui indiquent un effet du service fourni sur l'utilisateur, qui a un effet sur son comportement envers ledit service. Dans cette perspective, nous utilisons les résultats de l'UX afin de personnaliser les objets recommandés via notre système proposé.

2. Recommandation basée sur l'apprentissage par renforcement

Notre système permet de faire des recommandations basées principalement sur l'expérience utilisateur, moyennant le RL. Les types de données impliquées dans ce système sont les suivants:

1) Le sujet UX est l'entité qui interagit à travers différentes activités avec les objets UX. Il peut s'agir d'un utilisateur unique ou d'une communauté utilisant le système, effectuant des tâches telles que la création, la recherche, la modification et la suppression de contenu, et interagissant les uns avec les autres. **2) L'intention** représente l'objectif qui guide les interactions des utilisateurs au sein du système. Elles sont intimement liées au spectre des services offerts, c'est-à-dire à l'objet UX, mais aussi

au comportement de l'utilisateur. **3) Le Persona** est une abstraction conceptuelle représentant le comportement archétypal de l'utilisateur. Les personas sont dérivés des caractéristiques combinées des utilisateurs, des ensembles de tâches critiques et des données démographiques. **4) L'expérience utilisateur** (UX) représente l'expérience des sujets UX envers un objet UX. Il s'agit d'un ensemble d'activités réalisées sur une période de temps définie.

Afin de pouvoir utiliser le Reinforcement Learning, nous décrivons dans cette sous-section l'environnement de l'agent RL utilisé. Comme le montre la figure 1, le modèle RL est principalement composé d'un agent RL et d'un environnement, qui interagissent entre eux à travers trois (3) aspects principaux qui sont les États, les Actions et les Récompenses du système.

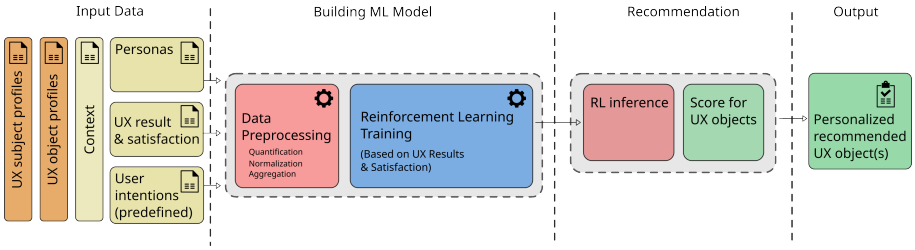


FIGURE 1 – Vue globale de la méthode proposée

1) Les états de l'agent du modèle RL incluent les caractéristiques des utilisateurs ou des sujets UX, les caractéristiques des objets UX, et les informations contextuelles. **2) Les Actions** de l'agent RL correspondent à la recommandation d'un objet UX spécifique de l'inventaire disponible. Ainsi, l'espace d'action est discret et représente l'ensemble de tous les objets UX de l'inventaire. L'objectif d'un tel agent RL est que pour un certain État, le système doit être capable de fournir des recommandations visant à guider l'utilisateur dans l'accomplissement de ses tâches. **3) Les récompenses** de l'agent RL sont basées sur les résultats UX obtenus, tels que les retours utilisateurs et le score de satisfaction. Nous fournissons à l'agent une agrégation des différents résultats de l'UX comme paramètre de récompense.

3. Conclusion

Nous proposons une recommandation basée sur le RL, et qui inclut le contexte, les intentions et les résultats UX. Cette approche peut être limitée en termes d'évolutivité, et ses performances sur des ensembles de données de différentes tailles restent à tester.

Bibliographie

Benramdane M. K., Kornyshova E. (2024). Reinforcement learning to personalize user experience within digital business ecosystems. In *2024 ieee 48th annual computers, software, and applications conference (compsac)*, p. 584–593.

Abstraction multiniveaux des graphes de connaissances : une approche de réification basée sur les graphes de propriétés

Selsebil Benelhaj-Sghaier[1], Annabelle Gillet[1] et Éric Leclercq[1]

1. Laboratoire d'Informatique de Bourgogne, Université Bourgogne Europe
9 avenue Alain Savary, F-21078 Dijon CEDEX, France
Selsebil_Ben-El-Haj-Sghaier@etu.u-bourgogne.fr,{prenom}.{nom}@u-bourgogne.fr

RÉFERENCE DE L'ARTICLE INTERNATIONAL Cet article est une synthèse de l'article : Benelhaj-Sghaier Selsebil, Gillet Annabelle, Leclercq Éric : Knowledge Graph Multilevel Abstraction : A Property Graph Reification Based Approach. RCIS 2024 : 12-19.

L'évolution continue de la diversité des données ainsi que la multiplication des cas d'usage révèlent de nouveaux défis en gestion des données et des connaissances, notamment sur la construction d'outils flexibles pour la modélisation et manipulation des données. Dans ce contexte, les graphes de connaissances (Hogan et al., 2021) permettent de structurer, stocker et interroger les données et les connaissances. Ils adoptent une approche basée sur les graphes, où les entités sont représentées par des nœuds et les relations par des arêtes. Les graphes de connaissances s'appuient sur différents modèles de graphes ayant des niveaux d'expressivité variés. Cependant, l'expressivité des modèles de graphes actuels reste encore assez limitée, ce qui nous encourage à étendre le modèle de graphe de propriétés (Angles, 2018) en intégrant un mécanisme de réification. Dans un modèle de graphe, la réification consiste à considérer un élément ou un ensemble d'éléments du graphe en tant qu'un autre élément du graphe. Par exemple, considérer un sous-graphe comme un nœud. Notre approche permet de représenter des relations complexes et de gérer plusieurs niveaux d'abstraction. Dans cet article, nous synthétisons les travaux que nous avons présentés dans (Benelhaj-Sghaier et al., 2024) qui détaillent un modèle de réification pour les graphes de propriétés.

Dans la première partie du travail nous avons étudié les différents modèles de graphes en montrant leurs limites et nous avons conclu que le modèle de graphe de propriétés couvre la plupart des fonctionnalités des autres modèles de graphes. Dans une seconde partie, nous avons étudié les techniques de réification et plus particulièrement celles pour RDF dont la limite majeure est qu'elles ne permettent pas de réifier plus qu'un triplet (deux nœuds connectés par une arête). Ainsi, toutes les approches de réification RDF ont été uniquement appliquées sur le modèle de

graphe orienté étiqueté. Donc, nous nous concentrons sur l'extension du modèle de graphe de propriétés en intégrant un mécanisme de réification qui permet de transformer un sous-graphe en un nœud réifié afin de pouvoir représenter des relations complexes et de définir des niveaux d'abstraction multiples. Ceci permet d'étendre le modèle de graphe de propriété existant (Angles, 2018) de la manière suivante :
$G = (V, E, R, \rho, \lambda, \sigma, \alpha)$ où $R \subseteq V$, R est un sous-ensemble de V contenant les nœuds réifiés et $\alpha: R \rightarrow SG$ une fonction totale qui fait correspondre les nœuds réifiés à leurs sous-graphes. Si $r \in R$ et $sg \in SG$, alors $\alpha(r) = sg$ où sg est le sous-graphe du nœud réifié r. SG est défini par un ensemble fini de tuples sg de la forme suivante : $sg = (V_r, E_r, R_r, \rho_r, \lambda_r, \sigma_r, \alpha_r)$ où $V_r \subseteq V$, $E_r \subseteq E$, $R_r \subseteq R$, les fonctions ρ_r, λ_r, σ_r et α_r sont les restrictions des fonctions ρ, λ, σ et α qui associent respectivement à une arête orientée ses nœuds, à un nœud ou une arête un ensemble d'étiquettes, à un nœud ou une arête ses propriétés et leur valeur, et à un nœud réifié son sous-graphe. Le mécanisme de réification récursive permet d'intégrer dans un nœud réifié $r \in R$ d'autres nœuds réifiés appartenant à l'ensemble R_r. Pour construire un nœud réifié r, il faut passer en paramètre de la fonction de construction β uniquement les éléments concernés par la réification de la manière suivante : $\beta(V_r, E_r, \lambda_r, \sigma_r) = r$. La fonction β respecte bien la réification récursive car V_r peut contenir des nœuds réifiés qui constituent le nœud réifié r.

Afin de pouvoir augmenter l'expressivité des graphes de connaissances, nous avons proposé un modèle de graphe de propriétés réifié qui permet d'abstraire un sous-graphe sous forme de nœud se comportant comme un nœud standard du graphe ce qui permet de représenter des relations complexes. Ainsi, grâce à la récursivité, il devient possible de définir plusieurs niveaux d'abstraction et de leur associer des propriétés, des labels ou des liens vers d'autres nœuds du graphe (réifiés ou non).. L'amélioration de l'expressivité des graphes de connaissances ne s'arrête pas jusqu'à ces travaux, nous avons étendu le modèle proposé en ajoutant un autre mécanisme de réification permettant d'abstraire un sous-graphe sous forme d'une arête réifiée. Par la suite, nous avons élaboré des exemples comparatifs avec les modèles classiques de réification et nous avons expliqué la sémantique du modèle proposé ainsi que ses opérateurs en Datalog. Dans la poursuite de nos travaux, nous travaillerons sur l'implémentation des opérateurs d'interrogation du modèle.

Bibliographie

Angles R. (2018). The property graph database model. In AMW, vol. 2100, p. 1–8.

Benelhaj-Sghaier S., Gillet A., Leclercq É. (2024). Knowledge graph multilevel abstraction: A property graph reification based approach, RCIS (pp. 12–19).

Hogan A., Blomqvist E., Cochez M., d'Amato C., Melo G. et al. (2021). Knowledge graphs, ACM Computing Surveys (CSUR), vol. 54, n°4, p. 1–37.

Intégration des dépendances fonctionnelles dans la définition de schéma des graphes de propriétés

Maude Manouvrier[1], Khalid Belhajjame[1]

1. Université Paris-Dauphine, PSL Research University
CNRS UMR [7243] LAMSADE
Place du Maréchal de Lattre de Tassigny 75775 Paris cedex 16, France
prenom.nom@dauphine.fr

REFERENCE DE L'ARTICLE INTERNATIONAL
Cet article est une synthèse de l'article : Manouvrier, M., Belhajjame, K. (2024). PG-FD: Mapping Functional Dependencies to the Future Property Graph Schema Standard. In Advances in Databases and Information Systems. ADBIS 2024. LNCS vol 14918. Springer, Cham, pp. 45-59. https://doi.org/10.1007/978-3-031-70626-4_4

Un graphe de propriétés est composé de nœuds et d'arêtes, associés à une étiquette et décrits par un ensemble de propriétés ou d'attributs, généralement représentés par des couples clé-valeur. Les graphes de propriétés sont largement utilisés dans de nombreux domaines. En avril 2024, le langage GQL, pour *Graph Query Language* (cf. Francis et al., 2023), a été officiellement publié en tant que norme ISO/CEI pour interroger les bases de données graphes. Angles et al. (2023) ont proposé d'étendre la norme GQL en définissant un formalisme, nommé *PG-Schema,* pour spécifier les schémas des graphes de propriétés, au sens classique du schéma dans les bases de données relationnelles (Barret et al., 2024). *PG-Schema* permet de définir des types de nœuds et d'arêtes, ainsi que des contraintes d'intégrité.

Les dépendances fonctionnelles (cf. Codd, 1972) sont des contraintes d'intégrité particulières. En base de données relationnelles, une dépendance fonctionnelle (DF) se définit par $X \rightarrow Y$, où X et Y sont des ensembles d'attributs d'un schéma R, et signifie que pour toutes les instances r de schéma R et pour tous les nuplets t_1 et t_2 de r, si t_1 et t_2 ont la même valeur pour X alors ils ont la même valeur pour Y. Les DFs sont notamment utilisées pour déterminer les identificateurs ou clés, contrôler la cohérence des données, optimiser les requêtes ou nettoyer des données. Plusieurs approches ont défini des dépendances fonctionnelles pour les bases de données graphes. Parmi ces approches, on peut citer GED (*Graph Entity Dependency*) de Fan et Lu (2019), gFD (*Graph-tailored functional dependency*) de Skavantzos et Link

(2023) et GD (*Graph Dependency*) de Zheng et al. (2023). GED et GD se basent sur des motifs de graphe (*Graph Pattern*), c'est-à-dire des sous-graphes orientés, dont les nœuds sont associés à des variables, permettant de spécifier le champ d'application de la DF. L'approche gFD se base, quant à elle, sur des conditions d'existence. Ces approches offrent une base théorique pour définir ou vérifier les DFs dans les bases de données graphes mais ne définissent pas de langage pour exprimer ces dépendances. L'objectif de notre proposition est de pallier ce manque et de spécifier comment traduire ces dépendances dans la future norme de définition de schéma de graphes, *PG-Schema*, de Angles et al. (2023).

Notre article présente une double contribution. Premièrement, il fournit une synthèse des approches de la littérature définissant des dépendances fonctionnelles pour les graphes, en soulignant notamment comment ces propositions sont liées les unes aux autres. Il présente également des règles de traduction des DFs pour les graphes en *PG-Schema* et démontre que ces règles respectent trois propriétés importantes, à savoir la calculabilité, la préservation de l'information et la préservation de la sémantique. Ces règles de correspondance sont mises en œuvre dans un prototype développé en Python et nommé PG-FD. Notre approche est, à notre connaissance, la première solution capable de transformer les dépendances fonctionnelles de graphes dans le standard *PG-Schema*, tout en préservant leur sémantique. Pour la suite de nos travaux, nous souhaitons mener des expérimentations plus poussées de notre prototype en utilisant des graphes réels et prendre en compte d'autres types de contraintes dans les graphes de propriétés.

Bibliographie

Angles, R., Bonifati, A., Dumbrava, S., *et al.* (2023). PG-Schema: Schemas for property graphs, *ACM on Management of Data*, vol. 1, n° 2, p. 1-25.

Barret, N., Enache, T., Manolescu, I., *et al.* (2024). Finding the PG schema of any (semi) structured dataset: a tale of graphs and abstraction. In *Proceedings of IEEE 40th International Conf. on Data Engineering Workshops (ICDEW), Utrecht, Netherlands*, pp. 365-369.

Codd, E.F (1972). Further normalization of the data base relational model. *Data base systems*, vol. 6, p. 33–64

Fan, W., Lu, P. (2019). Dependencies for graphs. *ACM Transactions on Database Systems (TODS)*, vol. 44, n° 2, p. 1-40.

Francis, N., Gheerbrant, A., Guagliardo, P., *et al*. (2023). A Researcher's Digest of GQL. In *Proceedings of 6th International Conf. on Database Theory (ICDT), Ioannina, Greece*. pp. 1:1-1:2.

Skavantzos, P., Link, S. (2023). Normalizing Property Graphs. *VLDB Endowment*, vol. 16, n° 11, p.3031–3043.

Zheng, X., Dasgupta, S., Gupta, A. (2023). P2KG: Declarative construction and quality evaluation of knowledge graph from polystores. In: *Proceedings of the 27th European Conf. on Advances in Databases and Information Systems (ADBIS), Barcelona, Spain*, pp. 427-439.

Approche non-supervisée pour la création d'un Référentiel Sémantique

Lydia Khelifa Chibout [1], Manuele Kirsch Pinheiro[2]

1. Centre Scientifique et Technique du Bâtiment (CSTB)
Lydia.CHIBOUT@cstb.fr
2. Centre de Recherche en Informatique, Université Paris 1 Panthéon-Sorbonne
Manuele.Kirsch-Pinheiro@univ-paris1.fr

RESUME. Les organisations font face à des défis sans précédent en matière de gestion et de structuration de leurs connaissances. La capacité à extraire, organiser et utiliser des informations pertinentes à partir de vastes collections de documents est devenue un facteur clé pour l'efficacité opérationnelle et la prise de décision éclairée. Cependant, l'identification des sources de connaissances nécessaires et la construction de bases de connaissances appropriées représentent une tâche ardue et chronophage. Cet article aborde ces défis en exploitant des techniques de traitement du langage naturel (NLP) et des modèles de langage de grande taille (LLM), afin de faciliter la création et l'enrichissement de vocabulaires spécialisés à des fins de gestion des connaissances. Nous explorons l'application de techniques non supervisées de clustering combinées à l'extraction de mots-clés avec des techniques de NLP pour l'aide à la construction de vocabulaires spécialisés répondant à la nature multidisciplinaire du CSTB, centre de recherche scientifique français spécialisé dans le bâtiment. Nous détaillons ici l'approche proposée, les résultats de nos expérimentations au CSTB, ainsi que le processus de validation humaine utilisé pour évaluer ces résultats.

Mots-clés : Extraction de mots-clés, clustering, identification du vocabulaire, construction basée sur les connaissances, gestion des connaissances.

ABSTRACT. The exponential growth of digital information has exposed organizations to unprecedented challenges in managing and structuring their knowledge repositories. The ability to extract, organize, and use relevant information from large collections of documents has become a critical factor for operational efficiency and informed decision-making. However, identifying necessary knowledge sources and building appropriate knowledge bases represents a cumbersome and time-consuming task. In this paper, we address these challenges by leveraging advanced Natural Language Processing (NLP) techniques and Large Language Models (LLMs), to facilitate the creation and enrichment of vocabularies for knowledge management purposes. We explore the application of clustering techniques combined with NLP-driven keyword extraction to support the construction of specialized vocabularies that address the multidisciplinary nature of the content at CSTB, a French scientific research center specialized on buildings. We provide a detailed overview of the proposed approach, present the results of our experiments, and describe the human validation process used to evaluate these results.

KEYWORDS: keywords extraction, clustering, vocabulary identification, knowledge-based construction, knowledge management

1. Introduction

La gestion de connaissances est devenue au fil des années un élément clé pour les organisations. Celle-ci, comme l'ensemble de notre société, est caractérisée par une certaine infobésité, avec la production toujours croissante de documents et des données. Dans ce contexte, la construction de bases documentaires efficaces au sein des organisations est devenue une tâche cruciale pour une bonne gestion de connaissances. En mettant en place des systèmes bien définis pour stocker, récupérer et indexer les documents, les organisations peuvent s'assurer que les informations pertinentes sont facilement disponibles pour ceux qui en ont besoin, permettant ainsi une prise de décision rapide et éclairée (Maharjan, 2020), (Morse, 2000). Cet accès simplifié à l'information facilite le partage des connaissances et la collaboration, favorisant un environnement dynamique propice à l'amélioration continue des processus (Maharjan, 2020), (Narazaki et al, 2020). Ceci est d'autant plus vrai dans un environnement multidisciplinaire, comme le CSTB, Centre Scientifique spécialisé dans le bâtiment, dont les projets couvrent des domaines multiples, tels que les sciences de l'environnement, la biodiversité, l'acoustique, les matériaux de construction et la santé dans les bâtiments.

Cependant, identifier les sources de ces connaissances nécessaires représente une tâche ardue et chronophage. Or la capacité à identifier facilement les informations pertinentes permet de prendre de meilleures décisions, de résoudre les problèmes plus efficacement et de contribuer plus efficacement aux objectifs organisationnels. Un vocabulaire bien structuré facilite non seulement la récupération des documents pertinents, mais améliore également l'efficacité globale des processus de gestion de l'information et de prise de décision. L'extraction de mots-clés est une étape clé de ce processus, car elle permet l'identification automatique des termes essentiels qui encapsulent le contenu principal des documents et aident les lecteurs à comprendre rapidement l'idée du contenu. En extrayant des mots-clés de plusieurs documents, il devient possible d'identifier des termes partagés entre plusieurs disciplines. Ces mots-clés communs agissent comme des ponts entre des domaines distincts, mettant en évidence des zones de chevauchement conceptuel ou d'intérêts partagés. Ce processus favorise la collaboration interdisciplinaire en fournissant un point de départ clair pour des projets et discussions conjoints. Par exemple, au sein du CSTB, l'identification de termes partagés tels que « durabilité » ou « modélisation des données » pourrait connecter les équipes de recherche en sciences de l'environnement et en informatique, permettant le développement de solutions innovantes interdisciplinaires. Cette approche rationalise l'intégration des connaissances et encourage la synergie entre les équipes multidisciplinaires.

Ces mots-clés directement liés au contexte du document constituent ainsi un référentiel sémantique. Pour y parvenir, les techniques d'extraction de mots-clés non supervisées ont attiré l'attention, en particulier dans les contextes où les ensembles de données étiquetés sont indisponibles ou impraticables à créer. Les récents progrès en traitement du langage naturel (NLP), y compris le développement de grands modèles de langage (LLM) tels que BERT (Devlin et al, 2019) et GPT (Brown et al, 2020), offrent des opportunités prometteuses pour améliorer la qualité et la précision

de l'extraction de mots-clés. Ces modèles, grâce à leur compréhension contextuelle approfondie, peuvent capturer des relations nuancées entre les mots et les concepts dans le texte, fournissant une base solide pour les approches non supervisées. Des études telles que celles (Mihalcea et Tarau, 2004) ou encore (Wan et Xiao, 2008) ont établi la base de l'extraction de mots-clés non supervisée. Plus récemment, BERTopic (Grootendorst, 2022) a émergé comme un outil efficace pour analyser et résumer de grands corpus, les rendant hautement pertinents pour la construction de référentiels sémantiques dans des environnements riches en connaissances.

Dans cet article, nous combinons les méthodes non supervisées avec des LLMs pour extraire des mots-clés d'un corpus documentaire, dans une démarche cohérente contribuant à la construction d'un vocabulaire spécialisé pour soutenir la récupération de documents et le partage des connaissances. L'objectif est de rendre la recherche d'informations plus précises et pertinentes. Le regroupement de documents est crucial pour les chercheurs engagés dans des recherches interdisciplinaires sur divers sujets. Notre approche propose l'extraction de mots-clés non-supervisée après le regroupement de documents textuels, ce qui améliore considérablement la découverte d'informations utiles et contribue à la compréhension et à la recherche d'information par les utilisateurs. Nous illustrons notre approche sur un ensemble de documents bilingues et multidisciplinaires du CSTB, dont les mots-clés identifiés ont été soumis à un panel d'experts du domaine à des fins d'évaluation.

Cet article est structuré comme suit : la section 2 présente les travaux connexes sur les méthodes d'extraction de mots-clés supervisées et non supervisées. La section 3 détaille nos approches proposées, suivie de la présentation des expérimentations réalisées au CSTB (section 4). La section 5 discute des résultats et du processus d'évaluation et de validation. Enfin, la section 6 conclut l'article.

2. Etat de l'art

L'organisation d'une base documentaire efficace présente de nombreux avantages aux organisations, quel que soit leur secteur d'activité (Laihonen et al., 2023 ; Jain, 2012 ; Yao-Sheng, 2007). Dans chacun de ces cas, un accès efficace aux connaissances explicites s'est traduit directement par une amélioration des performances, une innovation accrue et une efficacité globale augmentée. Pour cela, l'identification des mots-clés pertinentes représente une étape clé pour la construction de ces bases. Différents travaux dans la littérature traitent l'identification et l'extraction de mots-clés citons par exemples les méthodes traditionnelles qui reposent sur des analyses statistiques, telles que TF-IDF (Salton et Buckley, 1988), et linguistiques, comme la lemmatisation et l'extraction de syntagmes nominaux (Delamaire et al, 2019). Toutefois, elles peinent souvent à capturer la richesse sémantique des textes complexes. Des algorithmes non supervisés, tels que TopicRank (Bougouin et al., 2013), introduisent une approche par graphe pour structurer les mots-clés autour de thèmes cohérents ou encore Khelifa et al. (2012) qui proposent de structurer les mots en graphe de topics en gardant leurs contextualisations dans le texte à travers les dimensions sémantiques

telles que la région, le temps, la discipline/le domaine ou encore la langue. Abilhoa et De Castro (2014) proposent une méthode d'extraction de mots-clés pour les collections de tweets qui représente les textes sous forme de graphes et applique des mesures de centralité pour trouver les sommets pertinents (mots-clés). Hasan et al. (2018) proposent un système qui extrait un nombre spécifique de termes clés des documents pour identifier le contenu principal d'un texte. Les données sont collectées à partir de différentes sources telles que des livres et des journaux. Diverses techniques bien connues de l'apprentissage automatique, comme le SVM, la régression logistique ou l'arbre PAT-tree, ont été utilisées pour extraire les mots-clés. Bisht (2022) quant à lui a évalué différentes méthodes d'extraction de mots-clés basées sur la distribution spatiale et a proposé une mesure basée sur la fréquence, la fréquence inverse des documents, la variance et l'entropie de Tsallis, dont les résultats ont mis en évidence le fait qu'il n'existe pas de méthode parfaite. Ahadh et al. (2021) ont proposé une approche automatisée, semi-supervisée et indépendante du domaine pour analyser les rapports d'accidents. Étant donné un ensemble de sujets de classification définis par l'utilisateur et la littérature du domaine telle que des manuels, des glossaires et des articles Wikipédia, la méthode peut identifier des mots-clés spécifiques au domaine et les regrouper en sujets avec une implication minimale d'experts. Ces mots-clés et sujets peuvent ensuite être utilisés à diverses fins de fouille de données, y compris la classification. Cependant, ces méthodes nécessitent un nombre élevé de documents étiquetés comme exemples d'entraînement. Les approches plus récentes ont exploré l'utilisation des techniques d'apprentissage profond pour améliorer la précision et l'efficacité de l'extraction de mots-clés, démontrant le potentiel des réseaux neuronaux à apprendre des motifs complexes dans les textes et à identifier les mots-clés pertinents (Umair et al., 2024).

Par ailleurs, le clustering de documents, avec des techniques telles que K-means ou DBSCAN (Ester et al., 1996), permet d'améliorer la contextualisation des mots-clés en regroupant des documents similaires, mais l'évaluation de la qualité du clustering repose uniquement sur des indices comme le Silhouette Score (Rousseeuw, 1987). Les grands modèles de langage (LLMs) tels que BERT (Devlin et al., 2019) ou GPT-4 (OpenAI, 2023) offrent de grandes capacités pour comprendre le contexte sémantique profond, améliorant significativement la pertinence et la diversité des mots-clés extraits (Liu et al., 2019). Leur usage en classification multi-label (Tsoumakas & Katakis, 2007) optimise également la cohérence des résultats. Zhou et al. (2023) ont expérimenté l'utilisation de ChatGPT pour l'extraction de mots-clés, obtenant un ensemble représentatif et opérationnel pour la recherche scientifique.

On observe dans la littérature une tendance vers des approches supervisées, avec des mots-clés qui ne sont pas toujours regroupés en topic, et une confrontation au regard d'experts humains qui n'est pas toujours mise en avant. Or cette évaluation par des experts nous semble essentielle, notamment dans le cadre d'environnement multidisciplinaires, comme le CSTB. Combiner clustering et LLMs optimise la pertinence des mots-clés et réduit le bruit dans les données. Cette approche est particulièrement efficace pour des tâches telles que l'indexation intelligente, la

recherche d'information (Manning et al., 2008), la veille technologique et la recommandation de contenu personnalisée.

Dans cet article, nous proposons une approche non-supervisée permettant l'analyse d'un large volume de documents non étiqueté, combinant les approches de clustering de documents et d'extraction de mots de clés avec les LLM dont les résultats ont été présentés à un panel d'experts pour évaluation et validation.

3. Contribution

Dans cet article, nous proposons d'utiliser des techniques de traitement du langage naturel (TAL) pour extraire des mots-clés significatifs à partir de documents textuels et de les regrouper en topic. Le processus suit une chaîne structurée conçue pour optimiser les tâches de modélisation de sujets et de regroupement. Cette approche proposée représente une méthode non supervisée et modulaire capable de gérer un large corpus de documents multidisciplinaires non étiquetés. Un aperçu de ces différentes étapes du pipeline est présenté dans la Figure 1 :

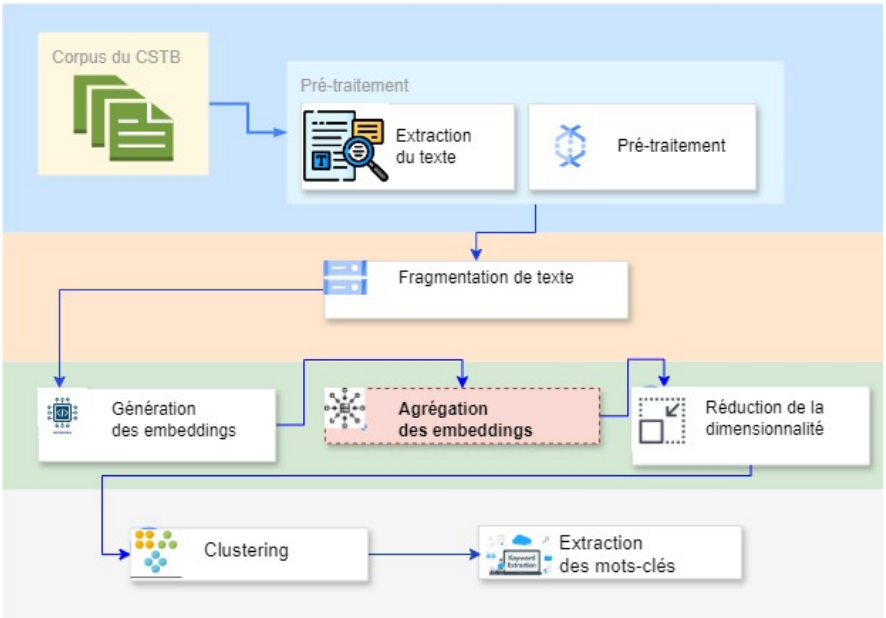


Figure 1. Approches de clustering de documents et de chunks

1. Extraction du texte et pré-traitement du corpus documentaire ;
2. Fragmentation de texte (*text chunking*) : les documents d'entrée sont divisés en segments plus petits et cohérents appelés *chunks* pour faciliter leur représentation et de permettre une analyse plus précise ;

3. Génération *d'embeddings* et réduction de la dimensionnalité : des modèles basés sur des *transformers* sont utilisés pour convertir les segments de texte (*chunks*) en vecteurs qui encapsulent la signification sémantique et pour réduire la dimensionnalité des *embeddings* tout en préservant leur structure ;
4. L'agrégation des *embeddings* est une étape obligatoire pour l'approche de clustering de documents. Eependant elle n'est pas utilisée dans le clustering par *chunks*. En effet, deux approches ont été envisagées pour le clustering, une approche uniquement basée sur les documents et une basée sur les *chunks*, offrant toutes les deux des perspectives différentes sur les documents analysés ;
5. Réduction de dimensionalité permet de réduire le nombre de dimensions (variables) tout en gardant les caractéristiques principales de chaque *embedding* ;
6. Clustering : les *embeddings* dont la dimensionalité a été réduite sont regroupés, pour identifier efficacement les régions denses et séparer le bruit ;
7. Extraction de mots-clés extrait des termes/mots clés et pour améliorer l'interprétabilité des sujets à partir des clusters en utilisant des techniques telles que c-TF-IDF et CountVectorizer.

Ces différentes étapes sont expliquées dans ce qui suit. L'ensemble de l'approche a été expérimentée sur une grande base de documents scientifiques et techniques du CSTB concernant le bâtiment, le génie civil, la qualité de l'air dans les bâtiments, les expérimentations acoustiques et d'autres domaines liés à la construction. Ces documents sont disponibles pour les chercheurs et les experts du CSTB, qui ont pu évaluer les mots-clés suggérés par l'approche proposée.

3.1. Extraction de texte et pré-traitement du corpus

La phase de prétraitement est une étape nécessaire pour n'importe quel processus d'analyse de données. Dans notre cas, celle-ci se traduit par l'extraction de texte à partir de documents PDF et leur prétraitement. Le format PDF est largement utilisé pour le partage de documents en raison de la présentation cohérente qu'il offre sur différents appareils. Cependant ce format présente des défis pour l'extraction automatisée de texte. En effet, les documents PDF ne sont pas conçus pour la manipulation de texte, et le contenu peut varier considérablement en complexité, incluant du texte brut, des tableaux, des images, et même des documents scannés contenant du texte sous forme d'images. Le prétraitement du texte est ainsi effectué afin de s'assurer que les données sont propres et prêtes pour une analyse ultérieure. Ce prétraitement consiste d'abord au *nettoyage* et à la *normalisation* du texte. Les caractères spéciaux tels que les guillemets typographiques et les apostrophes sont remplacés par leurs équivalents plus simples (par exemple, les guillemets courbes remplacés par des guillemets droits). De plus, les caractères de nouvelle ligne ***\n*** sont supprimés, et les espaces excessifs sont gérés en séparant et en concaténant les mots. Ensuite, tous les caractères numériques sont supprimés à l'aide d'expressions régulières. Par ailleurs, afin de réduire le bruit du texte et de focaliser l'analyse sur les éléments potentiellement significatifs, la *lemmatisation* et l'*étiquetage des parties*

du discours (POS) sont appliqués directement aux mots-clés et à l'extraction de bi-grammes. La *lemmatisation* permet de réduire les mots à leur forme de base ou racine. Par exemple, "running" devient "run" et "better" devient "good". Cette étape est cruciale pour s'assurer que les différentes formes d'un mot sont traitées pareillement lors de l'analyse. L'étiquetage des parties du discours, quant à elle, est utilisée pour identifier les verbes dans les mots-clés extraits, qui sont ensuite exclus pour se concentrer uniquement sur les noms et les adjectifs qui sont plus susceptibles de contribuer à la modélisation des sujets et à l'extraction des termes clés. Enfin, l'extraction de bi-grammes est réalisée.

3.2. Processus de segmentation du document

La segmentation du document correspond au processus de division d'un grand corpus de texte en segments plus petits et plus gérables, communément appelés *chunks*. L'objectif principal est de s'assurer que chaque fragment reste dans les limites d'entrée d'un modèle (par exemple, les limites de *tokens* dans les modèles de langage), tout en préservant suffisamment de contexte pour maintenir sa signification. La méthode utilisée ici est la division de texte basée sur les caractères, dans laquelle un long texte est divisé en fragments en fonction d'un nombre spécifique de caractères. Afin d'améliorer la continuité entre les fragments, un chevauchement de texte (*text overlapping*) a été introduit. Celui-ci implique d'inclure une petite portion du fragment précédent au début du fragment suivant, permettant ainsi une meilleure rétention du contexte à travers les fragments.

3.2 Vectorisation et réduction de dimensionnalité

La vectorisation, ou encore *embedding,* est le processus de conversion du texte en vecteurs numériques qui capturent sa signification sémantique. Ces vecteurs forment les modèles sur lesquels les techniques de clustering sont appliquées. Les méthodes principales pour créer des *embeddings* sont le Word2Vec (Churh, 2017), BERT (Feng et al, 2020) et Sentence-BERT. Word2Vec, développé par Google, utilise deux approches principales : le Continuous Bag of Words (CBOW), qui prédit un mot basé sur son contexte, et le Skip-gram, qui prédit les mots de contexte à partir d'un mot donné. Ces modèles sont entraînés sur de grands corpus de texte, produisant des *embeddings* qui capturent les relations entre les mots. BERT (Bidirectional Encoder Representations from Transformers) est un modèle basé sur les transformateurs qui génère des *embeddings* sensibles au contexte. Pour les tâches nécessitant des *embeddings* au niveau des phrases ou des documents, des méthodes comme Sentence-BERT (SBERT) ou Doc2Vec sont utilisées. SBERT affine BERT pour produire des *embeddings* de phrases sémantiquement significatives, comparables par similarité cosinus. Ces techniques ont diverses applications en traitement du langage naturel, dont la recherche de similarité, le clustering de documents, et la récupération d'information.

La réduction de dimensionnalité est un processus clé en apprentissage automatique, visant à réduire le nombre de caractéristiques (ou dimensions) d'un jeu

de données tout en conservant un maximum d'informations pertinentes (Anowar et al., 2021). Plusieurs algorithmes sont couramment utilisés à cet effet, dont UMAP (Uniform Manifold Approximation and Projection), particulièrement adapté à la visualisation de données à haute dimension. Il préserve davantage la structure globale des données par rapport à t-SNE, ce qui le rend efficace aussi bien pour la visualisation que pour l'apprentissage du manifold (Leland et al., 2018). T-SNE (t-Distributed Stochastic Neighbor Embedding) est une autre technique non linéaire très populaire pour visualiser des données complexes. Elle repose sur la conversion des similarités entre les données en probabilités conjointes et cherche à minimiser la divergence de Kullback-Leibler entre les espaces de haute et de basse dimension (Van Der Matteen and Hinton, 2018). Enfin, le PCA (Principal Component Analysis) est une méthode linéaire de réduction de dimensionnalité, qui projette les données sur les directions maximisant la variance. Elle est largement utilisée pour réduire la complexité des ensembles de données tout en préservant le plus possible la variabilité (Abdi et Williams, 2010). Les résultats de ces processus sont fortement influencés par la langue des documents. Mélanger des documents en différentes langues peut avoir un impact négatif sur ces résultats. Cependant, les documents dans différentes langues sont, à nous jours, couramment présents dans les organisations, ce qui doit être pris en compte lors des étapes d'*embedding* et de la réduction de la dimensionnalité. Ainsi, lors des expérimentations réalisées au CSTB, l'ensemble de documents utilisé a pu être divisé en deux sous-ensembles, composés respectivement de documents en anglais et en français. Un modèle distinct adapté à chaque langue a été utilisé pour chacun des deux sous-ensembles.

3.3 Les approches de clustering

Au cours de la phase de *clustering*, nous avons évalué le type de clustering le plus approprié, en tenant compte du fait que les documents sont multidisciplinaires et qu'un seul article peut aborder des questions transversales couvrant plusieurs domaines. Afin d'identifier l'approche optimale à notre expérimentation, nous avons exploré deux méthodes de clustering : une basée sur les documents et une autre basée sur les *chunks*. Ces deux méthodes ont été testées pour déterminer la stratégie la plus efficace pour l'extraction de mots-clés à intégrer dans le vocabulaire.

3.3.1 Approche de clustering de document avec les LLM

L'approche de clustering par document met l'accent sur le clustering de documents entiers plutôt que de segments individuels. Pour ce faire, une méthode d'agrégation est utilisée pour fusionner les *embeddings* des segments appartenant à un même document en une représentation unifiée. Cet *embedding* agrégé est ensuite utilisé pour le regroupement, permettant d'identifier les similitudes entre les documents dans leur ensemble. Différentes méthodes d'agrégation peuvent être employées pour combiner les *embeddings* individuels des chunks en une seule représentation pour chaque document, telles que l'agrégation par moyenne ou celle par somme. La méthode d'agrégation par moyenne calcule la moyenne des *embeddings* de tous les chunks au sein d'un document. Chaque *embedding* de chunks

est un vecteur de grande dimension, et l'agrégation par moyenne crée un vecteur unique en moyennant les *embeddings*. Cette approche aide à créer une représentation équilibrée du document, dans laquelle chaque chunks contribue de manière égale à *l'embedding* final du document. Par ailleurs, l'agrégation par la somme combine les *embeddings* en additionnant les vecteurs de tous les segments au sein d'un document. Les deux méthodes permettent d'agréger les *embeddings* au niveau des segments en une représentation au niveau du document qui capture la signification globale de celui-ci.

3.3.2 Approche de clustering de chunks

L'approche de clustering par *chunk* consiste à regrouper de plus petits segments (*chunks*) d'un document plutôt que le document entier. En divisant le document en fragments et en les regroupant individuellement, cette méthode capture des motifs plus localisés et des nuances thématiques au sein du texte.

3.4 Extraction des mots-clés et des topics

Pour construire un vocabulaire spécialisé dans une organisation multidisciplinaire telle que le CSTB, nous avons besoin de mots-clés qui caractérisent efficacement les documents, facilitant ainsi l'accès aux connaissances qu'ils contiennent. Cette extraction de mots-clés et de sujets (topics) se déroule en deux étapes :

- **Vectorisation** : le texte de chaque document est transformé en une matrice de fréquence des termes (par document). Chaque entrée de cette matrice indique combien de fois un mot spécifique apparaît dans chaque document ;
- **Calcul du c-TF-IDF** (ou TF-IDF basé sur les classes) : une fois les clusters formés, le c-TF-IDF (Xu & Wu, 2014) est calculé pour chaque cluster. Cela fournit une pondération des termes basée sur leur importance relative au sein de ce cluster.

4. Expérimentation

Nous avons testé notre approche sur un large corpus de documents multidisciplinaires non étiquetés qui a été divisé en deux parties en fonction de la langue (français et anglais). Des modèles spécifiques pour les *embeddings* et les techniques de traitement du langage naturel ont été appliqués à chaque partie. L'objectif de cette expérimentation est triple : i) effectuer un clustering non supervisé de documents et de chunks ; ii) extraire des mots-clés en suivant les deux approches de clustering ; et iii) évaluer les résultats pour identifier la meilleure approche, les mots-clés et les sujets pour la construction du vocabulaire du CSTB.

4.1 Description du corpus documentaire

La plupart des documents disponibles au CSTB sont des rapports de recherche et d'évaluation publiés en format PDF entre 2000 et 2024, couvrant différents domaines scientifiques tels que le génie civil, la sécurité incendie, la santé et la qualité de l'air dans les bâtiments. Le choix de la période et des documents réside dans le fait que les documentalistes chargés de leur indexation au CSTB, et qui sont mobilisés dans le cadre de la validation des résultats, ont une très bonne connaissance de ce corpus (ces sont eux qui nous ont transmis ce corpus).

Ce corpus contient un total de 6627 documents, dont 3279 en français, 3055 en anglais et 293 documents jugés inexploitables. Les causes empêchant l'exploitabilité de ces documents sont nombreuses : documents scannés sans traitement OCR, rendant leur contenu non consultable ; des documents sont mal encodés, causant des problèmes de lisibilité et d'accessibilité ; et des PDF complètement vides, n'offrant aucune donnée utilisable. Pour les documents en français, l'analyse a identifié 2 808 fichiers uniques et 419 groupes de doublons, où chaque groupe contient des fichiers identiques réduits à un seul représentant. De même, pour les documents en anglais, il y a 2265 fichiers uniques et 716 groupes de doublons. Cette catégorisation détaillée et l'identification de doublons améliorent l'efficacité de la gestion des documents et aident à rationaliser l'analyse ultérieure en s'assurant que les données redondantes n'encombrent pas l'ensemble de données. Les résultats soulignent l'importance du prétraitement et de l'amélioration de la qualité des documents afin de garantir une meilleure utilisabilité dans la gestion des connaissances.

4.2 4.2 Modèles et techniques utilisés

L'expérimentation a été faite sur un ordinateur équipé d'un processeur Intel i5 de 5^{e} génération et de 16 Go de RAM DDR4 combiné à l'environnement de développement virtuel gratuit Google Colab pour son type d'exécution GPU.

Pour des contraintes techniques liées à l'infrastructure, notre choix s'est porté sur le modèle d'*embedding all-MiniLM-L6-v2* (Wenhui et al., 2020) et paraphrase-multilingual-MiniLM-L12-v2 (Ciancone et al., 2024), des modèles d'*embedding* par LLM qui appartiennent à la famille des modèles MiniLM, qui sont des versions allégées de BERT. Le MiniLM est conçu pour offrir des *embeddings* de haute qualité tout en étant plus léger et plus rapide que les modèles BERT. Concernant la réduction de la dimensionnalité, nous avons opté pour *UMAP* (McInnes et al., 2018), qui est rapide et préserve à la fois les structures de données locales et globales. Le choix de cette méthode est fortement lié au modèle de clustering HDBSCAN. Notre choix s'est porté sur ce modèle car il ne nécessite pas de spécifier le nombre de clusters à l'avance et gère bien les densités variables (Malzer & Baum, 2020). De plus, connaître la hiérarchisation des sujets sera très utile pour l'intégration des résultats dans le référentiel sémantique. L'extraction de mots-clés quant à elle, a été réalisée en utilisant *CountVectorizer* combiné avec *c-TF-IDF* (Xu & Wu, 2014), mettant en avant l'importance spécifique des clusters. En considérant l'approche

basée sur le clustering, nous avons appliqué une agrégation par moyenne sur les deux langues afin de fusionner les segments de chaque document.

5. Résultats and Evaluation

5.1 Processus d'évaluation et de validation

Compte tenu de l'importance des mots-clés extraits (candidats au vocabulaire spécialisé) et du rôle central de ce référentiel dans la stratégie de gestion des connaissances du CSTB, une validation humaine des résultats a été privilégiée. Ce processus d'évaluation s'appuie sur l'expertise métier et l'expérience approfondie des intervenants, mobilisés spécifiquement pour cette tâche.

Le panel d'experts se compose de : *(i)* Un ingénieur en génie civil qui a une double casquette et qui contribue à l'indexation documentaire (15 ans d'expérience au CSTB) ; *(ii)* Une chercheuse spécialisée en santé et confort des bâtiments (40 ans au CSTB) ; *(iii)* Deux documentalistes experts (28 et 20 ans d'expérience chacun en indexation de documents techniques et gestion des contenus) et occasionnellement un veilleur en technologique de l'information (14 ans d'expérience au CSTB).

Ce comité, déjà en charge de l'animation des ateliers de construction du référentiel sémantique du CSTB, a appliqué une approche d'analyse fondée sur : *(i)* La pertinence thématique des mots-clés pour leur intégration futur dans le référentiel sémantique ; *(ii)* Leur adéquation aux enjeux multidisciplinaires de l'organisme ; *(iii)* Leur potentiel d'interopérabilité avec les systèmes existants.

Les missions de ce comité sont, en fonction des résultats, de guider le choix des meilleures méthodes de clustering et des techniques de NLP, et la nomination des sujets ou topics correspondants aux clusters résultants pour identifier les domaines scientifiques. Le processus de validation est cyclique et itératif. Il s'est organisé en quatre réunions de travail (de 2 heures chacune). Au cours de celles-ci, les résultats du *clustering* et de l'extraction ont été présentés, et des exigences/recommandations ont été établies pour les mots-clés. Parmi celles-ci, le comité a établi 3 **critères de sélection** : *(i)* Les mots-clés ne doivent pas être des verbes ; *(ii)* Ils ne doivent pas contenir de chiffres, de noms de villes ou de pays ; *(iii)* Ils doivent se composer d'un-grammes ou de bi-grammes. Le comité a également défini des indications sur la **métrique d'évaluation** : Le nombre de documents par clusters pour décider d'intégrer ou non les mots-clés candidats au référentiel sémantique, permettant ainsi d'analyser la distribution des documents par thématique.

5.2 Résultat de l'approche de clustering de documents

Documents en anglais

Dans notre expérimentation menée sur 2265 documents en anglais, nous avons obtenus six clusters distincts. Chaque topic représente un cluster, caractérisé par les mots-clés extraits et le nombre de documents qu'il contient. Comme illustré dans la

Figure 2, nous pouvons voir qu'en utilisant l'extraction de bi-grammes, le topic numéro 3 contient les bi-grammes « *wind speed* » et « *wind tunnel* » que le comité de validation identifie comme des mots-clés importants, se référant à des expériences menées avec la soufflerie Jules Verne, une installation de recherche pour les évaluations techniques au CSTB. Le comité a trouvé que les bi-grammes fournissent des informations plus significatives par rapport aux simples mots-clés. Nous pouvons voir que la plupart des documents contiennent les mots « *building, constructions* » et des mots connexes, ce qui est normal compte tenu du contexte de tous les documents.

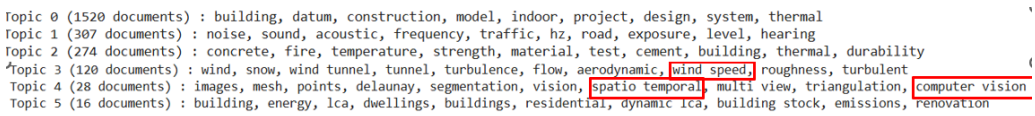


Figure 2. Résultats de l'extraction des mots bi-grammes avec l'approche de clustering des documents

Documents en français

A partir de 2808 documents en français, nous avons obtenus six clusters distincts. Comme pour l'expérimentation sur des documents en anglais, les mots bi-gramme ont été considérés comme plus pertinents et couvrant mieux certains domaines comme « la gestion patrimoine » et le « vide sanitaire ».

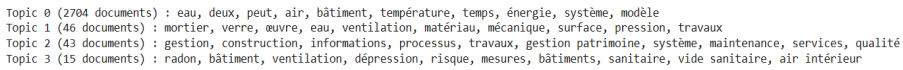


Figure 3: Résultats de l'extraction des mots bi-grammes avec l'approche de clustering des documents

5.3 Résultats de l'approche de clustering des chunks

Documents en anglais

Le même corpus de documents en anglais a été testé avec l'approche basée sur les *chunks*. Comme illustré dans la Figure 4, cette approche offre 90 clusters et plus de mots-clés. Les domaines sont plus largement couverts et plus étendus, garantissant qu'aucun domaine abordé par le CSTB n'a été négligé. Les bi-grammes extraits fournissent également une richesse sémantique accrue, selon les experts, similaire à la première approche. Le comité a en effet considéré que ces résultats étaient plus riches et plus proches de la nature du contenu des documents du CSTB.

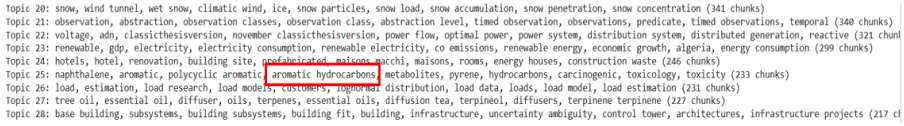


Figure 4. Résultats de l'extraction des mots bi-grammes avec l'approche de clustering des chunks

Document en français

Cependant, lors du test du corpus en français, nous avons rencontré de nombreux résultats incohérents pour les deux approches proposées, comme illustré dans la Figure 4. Par exemple, le sujet 10 dans la Figure 5 comporte plusieurs mots qui n'ont pas pu être interprétés par les experts, tels que « fissap » (« passif » en inversé), qui sont probablement une conséquence du chevauchement de texte utilisé lors de l'étape de fragmentation. Bien que d'autres sujets identifiés aient été considérés comme pertinents par les experts, ce phénomène démontre certaines limites de notre approche, indiquant qu'il reste encore des améliorations à apporter.

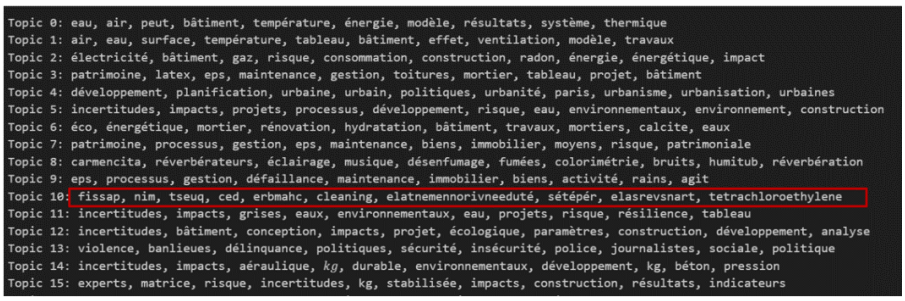


Figure 5. Résultats incohérents de l'extraction avec l'approche basée sur des chunks

5.5 Discussion

À l'issue des 4 réunions de travail, le comité a retenu les résultats de l'approche basée sur les *chunks* pour l'extraction des mots-clés, en mettant temporairement de côté le traitement des documents en français. Lors de la dernière réunion de validation, la méthode *chunk-based* a été désignée comme la solution optimale, en privilégiant l'utilisation de bi-grammes pour garantir une granularité sémantique adaptée aux besoins du projet. En effet, cette approche permet de capturer des unités de sens cohérentes (ex: « qualité de l'air », « changement climatique »), tandis que les bi-grammes évitent la sur-spécialisation des un-grammes isolés.

5 Conclusions et perspectives de recherche

Dans cet article, nous présentons les premiers résultats d'une approche non supervisée associant des grands modèles de langage (LLM) et des techniques de traitement automatique des langues (TAL). Cette combinaison permet d'extraire des mots-clés candidats pertinents, en vue de leur intégration ultérieure au sein d'un référentiel sémantique. L'approche proposée se décline en deux sous-approches de clustering : l'une fondée sur les *chunks* (segments thématiques) et l'autre sur les documents complets. L'expérimentation a été menée sur un corpus multidisciplinaire, bilingue et non étiqueté. Compte tenu de l'importance stratégique des mots-clés candidats et des directives de la politique globale de Gestion des Connaissances au CSTB, une validation humaine des résultats a été privilégiée. Le comité d'évaluation, composé d'experts aux spécialités, profils et ancienneté variés au sein du CSTB, a participé à quatre réunions. Lors de ces sessions, les exigences ont été définies, conduisant à un affinement des techniques de TAL et à une révision itérative des résultats pour en assurer la qualité. Les conclusions du comité soulignent que les résultats les plus performants proviennent de l'approche par *chunks*, offrant une couverture thématique tout en renforçant la cohérence des clusters. Alignés sur ses enjeux opérationnels, ls travaux futurs s'articuleront autour de trois axes :

1. L'optimisation du traitement des documents en français via des modèles linguistiques dédiés (ex. : CamemBERT) ;
2. La résolution des anomalies lexicales résiduelles (ex. : le terme « fissap », probable coquille pour « fissure ») ;
3. L'amélioration de la génération de mots-clés par intégration du *Maximum Marginal Relevance* (MMR), afin d'équilibrer pertinence et diversité. Enfin, l'attribution de libellés thématiques aux clusters (« Acoustique », « Énergétique »…) par le comité positionne ce référentiel comme un outil stratégique pour l'ingénierie multidisciplinaire, consolidant l'accès aux savoirs techniques du CSTB ;
4. Enrichissement du référentiel sémantique du CSTB avec ces mots clés extraits.

Bibliographie

Abilhoa W.D., De Castro L.N. (2014). TKG: A graph-based approach to extract keywords from tweets. *Distributed computing and artificial intelligence, 11th International Conference*, Springer International Publishing, p. 425-432.

Abdi, H., & Williams, L. J. (2010). Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 2(4), 433–459.

Anowar, F., Sadaoui, S., & Selim, B. (2021). Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP,

LE, ICA, t-SNE). Computer Science Review, 40, 100378. https://doi.org/10.1016/j.cosrev.2021.100378

Ahadh A., Binish G.V., Srinivasan R. (2021). Text mining of accident reports using semi-supervised keyword extraction and topic modeling. *Process Safety and Environmental Protection*, 155 (Nov. 2021), 455-465. https://doi.org/10.1016/j.psep.2021.09.022.

Bisht R.K. (2022). A Comparative Evaluation of Different Keyword Extraction Techniques. *International Journal of Information Retrieval Research (IJIRR)*, 12(1), 1-17.

Brown T.B., Mann B., Ryder N., Subbiah M., Kaplan J.D., Dhariwal P., Neelakantan A. et al. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901.

Carbonell J., Goldstein J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 335-336. ACM, Melbourne, Australia.

Ciancone M., Kerboua I., Schaeffer M., Siblini W. (2024). MTEB-French: Resources for French Sentence *Embedding* Evaluation and Analysis. *arXiv*, arXiv:2405.20468. Disponible sur : https://arxiv.org/abs/2405.20468.

Church Kw. Word2Vec. *Natural Language Engineering*. 2017; 23(1):155-162.

Devlin J., Chang M.-W., Lee K., Toutanova K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT 2019*, pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota, 2-7 juin 2019.

Delamaire, A., Beigbeder, M., & Juganaru-Mathieu, M. (2019, May). Exploitation de syntagmes dans la découverte de thèmes. Actes de la conférence CORIA (Conférence en Recherche d'Information et Applications).

Grootendorst M. (2022). BERTopic: Leveraging BERT and c-TF-IDF for Topic Modeling. *arXiv*, arXiv:2203.05794. Disponible sur : https://arxiv.org/abs/2203.05794.

Hasan H.M., Sanyal F., Chaki D. (2018). A novel approach to extract important keywords from documents applying latent semantic analysis. *2018 10th International Conference on Knowledge and Smart Technology (KST)*, pp. 117-122. IEEE, Chiang Mai, Thaïlande.

Khelifa LN., Lammari N., Akoka J., Bouabana-Tebibel T. (2012), *Building Contextualized Topic Maps*. 19th IBIMA (International Business Information Management Association) conference on Innovation Vision 2020: Sustainable growth, Entrepreneurship, Real Estate and Economic Development, Nov 2012, Barcelone, Spain. ⟨hal-01126211⟩

Laihonen H., Kork A.A., Sinervo L.M. (2023). Advancing public sector knowledge management: towards an understanding of knowledge formation in public administration. *Knowledge Management Research & Practice*, 22(3), 223–233. https://doi.org/10.1080/14778238.2023.2187719.

Leland McInnes, John Healy, and James Melville (2018). “UMAP: Uniform manifold approximation and projection for dimension reduction”. In: arXiv preprint arXiv:1802.03426

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2020). Language-agnostic BERT sentence embedding. arXiv preprint arXiv:2007.01852.

Maharjan P. (2020). Knowledge Management Enablers for Knowledge Creation Combination in Nepalese Hospitality Industry. *Journal of Balkumari College*, 9(1), 25-33. https://doi.org/10.3126/jbkc.v9i1.30064.

Malzer C., Baum M. (2020). A hybrid approach to hierarchical density-based cluster selection. *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, pp. 223-228. IEEE, Karlsruhe, Allemagne.

McInnes L., Healy J., Melville J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv*, arXiv:1802.03426. Disponible sur : https://arxiv.org/abs/1802.03426.

Mihalcea R., Tarau P. (2004). TextRank: Bringing Order into Texts. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 404-411.

Morse R. (2000). Management in the 21st Century Knowledge Management Systems: Using Technology to Enhance Organizational Learning. *Proceedings of the 2000 Information Resources Management Association International Conf. on Challenges of Information Technology Management in the 21st Century*, pp. 426–429. IGI Global, Anchorage, USA.

Narazaki R.Y., Silveira Chaves M., Drebes Pedron C. (2020). A project knowledge management framework grounded in design science research. *Knowledge and Process Management*, 27(3), 197-210. https://doi.org/10.1002/kpm.1627.

Priti J. (2012). An Empirical Study of Knowledge Management in University Libraries in SADC Countries. In : Hou H.T., *New Research on Knowledge Management Applications and Lesson Learned*. IntechOpen. https://doi.org/10.5772/36309.

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3982–3992). Hong Kong, China: Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Tsoumakas, Grigorios & Katakis, Ioannis. (2007). Multi-label classification: An overview. IJDWM. 3. 1-13.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, **9**, 2579–2605.

Wan X., Xiao J. (2008). Single Document Keyphrase Extraction Using Neighbourhood Knowledge. *Proceedings of the 23th AAAI Conference on Artificial Intelligence*, pp. 855-860. American Association for Artificial Intelligence, Chicago, Illinois, 13-17 juillet.

Wenhui W., Furu W., Li D., Hangbo B., Nan Y., Ming Z. (2020). MINILM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. *Proceedings of the 34th International Conf. on Neural Information Processing Systems (NIPS '20)*, Article 485, pp. 5776–5788. Curran Associates Inc., Red Hook, NY, USA.

Xu D.D., Wu S.B. (2014). An improved TFIDF algorithm in text classification. *Applied Mechanics and Materials*, 651 (Sep. 2014), 2258-2261.

Yao-Sheng L. (2007). The effects of knowledge management strategy and organization structure on innovation. *International Journal of Management*, 24(1), 53-60.

Zhou J., Jia Y., Qiu Y., Lin L. (2023). The potential of applying ChatGPT to extract keywords of medical literature in plastic surgery. *Aesthetic Surgery Journal*, 43(9), NP720-NP723.

Bonnes pratiques et mauvaises surprises de l'intelligence artificielle pour la gestion des connaissances tacites en entreprise

Pierre-Emmanuel Arduin [1], **Manuele Kirsch Pinheiro** [2], **Lydia Khelifa Chibout** [3]

1. *Université Paris-Dauphine, PSL, DRM UMR CNRS 7088*
   pierre-emmanuel.arduin@dauphine.psl.eu
2. *Université Paris 1 Panthéon-Sorbonne, CRI UR CNRS 1445*
   manuele.kirsch-pinheiro@univ-paris1.fr
3. *Centre Scientifique et Technique du Bâtiment*
   lydia.chibout@cstb.fr

RÉSUMÉ. *La transmission de connaissances tacites représente un important défi pour des nombreuses organisations. Ces mêmes organisations sont aujourd'hui confrontées à l'émergence de l'Intelligence Artificielle et les différentes transformations qu'elle apporte. Ces transformations interrogent et nous amènent à nous intéresser à l'apport que l'IA pourrait avoir dans la transmission des connaissances tacites. Cette question a animé l'atelier "Connaissances Tacites" réalisé dans le cadre de la conférence EGC 2025. Cet article résume les nombreuses échanges réalisées pendant l'atelier autour des bonnes pratiques et mauvaises surprises de l'IA pour la gestion de connaissances.*

ABSTRACT. *The transmission of tacit knowledge represents a major challenge for many organizations. These same organizations are now experiencing the emergence of Artificial Intelligence and the various transformations it brings on many domains. All these transformations lead us to consider how AI could contribute to the transmission of tacit knowledge. This question was the subject of the "Connaissances Tacites" workshop held as part of the EGC 2025 conference. This article summarizes the discussions held during the workshop on the good practices and bad surprises of AI for knowledge management.*

MOTS-CLÉS : *gestion de connaissances, connaissances tacites, intelligence artificielle*

KEYWORDS: *knowledge management, tacit knowledge, artificial intelligence*

1. Introduction

Les connaissances des entreprises reposent sur des savoirs et des savoir-faire, des compétences, des techniques et des informations. Ces connaissances peuvent être explicitées, c'est-à-dire formalisées dans des documents et des rapports, mais aussi tacites, reposant sur des processus interprétatifs et portées par des individus (Nonaka, Takeuchi, 1995). Ces dernières sont un ensemble de savoir-faire, d'intuitions et d'expériences personnelles difficiles à formaliser et à transmettre : « *we can know more than we can tell* » (Polanyi, 1967, p. 4). Elles se développent au fil du temps, avec l'expérience, au contact de situations concrètes, d'interactions sociales, en particulier dans un contexte donné (Kirsch-Pinheiro, 2023). Les connaissances tacites sont souvent inconscientes et profondément ancrées dans l'expérience individuelle. Elles sont essentielles dans de nombreux domaines, notamment dans les métiers créatifs, les professions manuelles ou encore la prise de décision stratégique. Malgré leur importance, les connaissances tacites restent un défi pour les organisations qui cherchent à les capturer, les partager et les pérenniser avec parfois des outils technologiques aux limites mal appréhendées (Arduin, Ziam, 2024).

Ces mêmes organisations ont vu l'émergence, depuis quelques années, d'applications de l'Intelligence Artificielle (IA) avec entre autres l'utilisation de graphes de connaissances (Mecharnia *et al*., 2021) dans différents domaines, comme le marketing, les ressources humaines, ou encore la finance. On peut alors s'interroger sur l'apport que l'IA, sous ses différentes formes, peut avoir en gestion des connaissances, en particulier concernant les connaissances tacites qui sont difficiles à formaliser et à transmettre, ainsi que sur les bonnes pratiques et retours d'expériences permettant de connaître les limites de l'IA pour gérer les connaissances tacites.

Cette question de l'apport de l'IA a été au centre de l'atelier "*Connaissances Tacites*" (KM-IA) [1] qui a eu lieu début 2025 dans le cadre de la Conférences EGC (*Extraction et Gestion des Connaissances*) [2]. L'objectif de cet article est ainsi de rapporter les discussions et les échanges ayant eu lieu pendant l'atelier entre chercheurs et industriels autour d'applications de l'IA en gestion des connaissances tacites, en abordant non seulement des bonnes pratiques, mais aussi des mauvaises surprises qu'il convient également de prendre en compte.

Le restant de cet article est organisé comme suit : La section 2 introduit les matériels et méthodes utilisées pour la réalisation de l'atelier et de cet article. La section 3 fait un rappel sur les concepts théoriques abordés, notamment les notions de connaissances tacites et d'Intelligence Artificielle. La section 4 résume les points les plus pertinents exposés par les différentes contributions présentées pendant l'atelier, alors que la section 5 présente les discussions qui ont eu lieu autour des apports et des limites de l'IA pour la gestion des connaissances tacites. Enfin la section 6 avance quelques conclusions et perspectives à partir des discussions établies.

1. https://km-ia.sciencesconf.org
2. https://www.egc2025.cnrs.fr

2. Matériels et méthodes

Le présent article propose une restitution des discussions et des échanges qui ont eu lieu à l'atelier « *Gestion des connaissances tacites en entreprise : réflexions, retours d'expériences, bonnes pratiques et mauvaises surprises de l'intelligence artificielle* » (KM-IA), proposé dans le cadre de la Conférence EGC 2025. Le sujet de cet atelier, l'apport et les limites de l'IA dans la gestion de connaissances, et plus particulièrement en ce qui concerne les connaissances tacites, touche particulièrement l'Informatique des Organisations et les Systèmes d'Informations, sujets centraux pour la communauté *Inforsid*, ce qui a motivé la réalisation de cet article.

La préparation de ce matériel, de la réalisation de l'atelier en lui-même, à la restitution, présentée dans la section 4, ainsi que son analyse, présentée dans la section 5, ont suivi les trois phases illustrées par la Fig. 1.

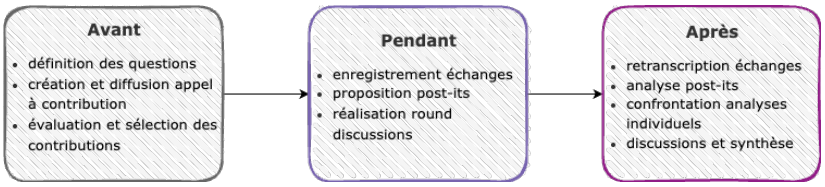


FIGURE 1. *Phase de travail, avant, pendant et après la réalisation de l'atelier*

Avant la réalisation de l'atelier, un certain nombre de questions ont été identifiées par les organisateurs comme étant pertinentes dans le but de susciter des discussions autour de la thématique de l'IA et les connaissances tacites. Parmi ces questions, nous avions :

– Comment l'IA peut aider les entreprises à partager et pérenniser les connaissances tacites portées par les individus au sein des organisations ?

– Quels usages des outils basés sur l'IA pour analyser les interactions et les communications afin d'en extraire des connaissances (analyse textuelle des comptes-rendus des réunions, des notes de réunions, des enregistrements vidéo de réunions, des données issues d'outils collaboratifs de l'entreprise, etc.) ?

– Vers quelles directions pointent les premiers retours d'expérience en entreprise ?

– Quelles sont les implications de l'application de l'IA sur l'organisation et ses employés ?

– Quelles limites managériales et éthiques émergent ?

Ces questions ont été reprises dans un appel à contributions qui a été diffusé par différents canaux de communications dans les communautés EGC et Inforsid. Après relecture par le comité de programme, huit contributions sur treize ont été retenues. Chaque contribution a été relue par trois membres du comité du programme, qui ont évalué la qualité scientifique et l'intérêt pour les discussions apporté par la contribution.

Pendant l'atelier, chaque contribution a été présentée suivie par un round de discussions libres avec l'ensemble des participants, composé d'une quinzaine des personnes, ayant ou non une contribution à l'atelier. Afin de proposer une restitution la plus fidèle possible, les présentations et les échanges ont été enregistrés.

Par ailleurs, pendant l'atelier, les participants avaient accès à huit affiches, chacun contenant la première page d'une contribution. Chaque participant a ainsi reçu un bloc des *post-it* et des stylos. Ces post-its avaient comme vocation de permettre à chaque participant de déposer des questions, des suggestions ou des commentaires sur une contribution, permettant ainsi aux participants de transmettre aux auteurs des questions ou des commentaires supplémentaires qu'ils n'auraient pas eu le temps d'aborder oralement pendant le temps d'échange après chaque présentation. La Fig. 2 illustre le résultat en fin de séance pour une de contributions. A la fin de l'atelier, ces post-its ont ainsi pu être documentés et récupérés pour analyse.

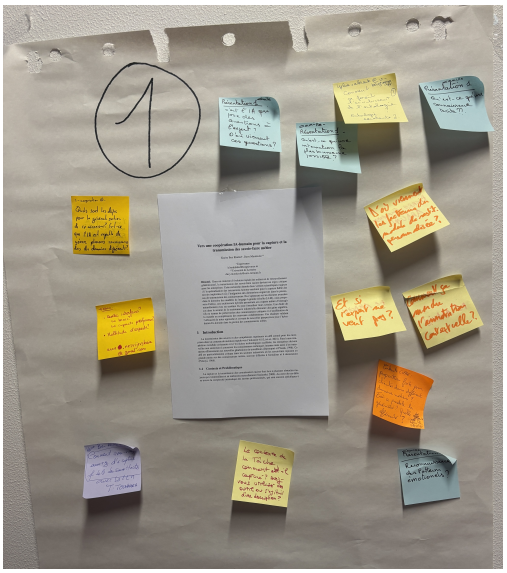


FIGURE 2. *Exemple de poster avec les post-its déposés par les participants*

Après la réalisation de l'atelier, l'ensemble d'échanges a été retranscrit et analysé, tout comme l'ensemble des post-its. Ces analyses ont été réalisées d'abord individuellement, puis nous avons confronté nos analyses individuelles pour en faire ressortir les points considérés comme les plus pertinents. Ceci a donné lieu à un round des discussions, qui a abouti à une synthèse et donc à l'ensemble de points présentés dans les sections 4 et 5.

Cependant, avant de présenter les résultats de cette synthèse et la restitution qui en découle, il nous parait opportun d'établir un certain nombre de concepts indispensables pour la bonne compréhension de l'ensemble des contributions.

3. Concepts théoriques

3.1. Les connaissances tacites, des processus cognitifs individuels

(Polanyi, 1967, p. 4) a été le premier à noter que nous savons plus que nous pouvons dire : « *We can know more than we can tell* ». Ce que nous savons mais ne pouvons dire, ce sont les connaissances tacites, qui sont difficilement exprimables, quelle que soit la forme du langage. Un matin de 1967, alors qu'il lisait son courrier, Polanyi lit une lettre et pensa qu'elle pourrait intéresser son fils. Il tendit la lettre à son fils, mais se souvint alors que son fils ne parlait qu'anglais, alors que la lettre n'était pas écrite en anglais. Polanyi réalise alors qu'il possédait le *sens* de la lettre mais pas le texte, c'est-à-dire les informations. C'est ce qu'il a appelé le *sense-reading* (Polanyi, 1967), processus cognitif central pour partager des connaissances tacites qui consiste en la lecture d'informations qui sont interprétées pour créer des connaissances tacites. Le *sense-giving* est le processus dual, par lequel les connaissances tacites qui sont portées par les individus sont explicitées sous forme de mots au travers de la création d'informations (Fig. 3).

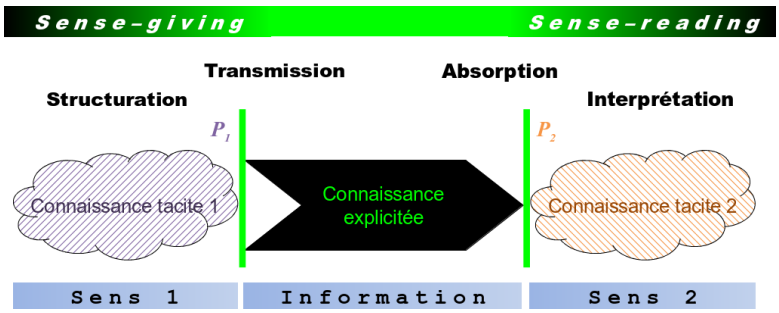


FIGURE 3. *Partage des connaissances tacites, sens et information (Arduin, 2018)*

Les individus, au travers de leurs schémas d'interprétation qui sont également appelés modèles mentaux, donnent un sens aux informations qu'ils créent à partir de leurs connaissances tacites et lisent un sens dans les informations qu'ils perçoivent pour créer leurs propres connaissances tacites. Ainsi dans les entreprises, on distingue :

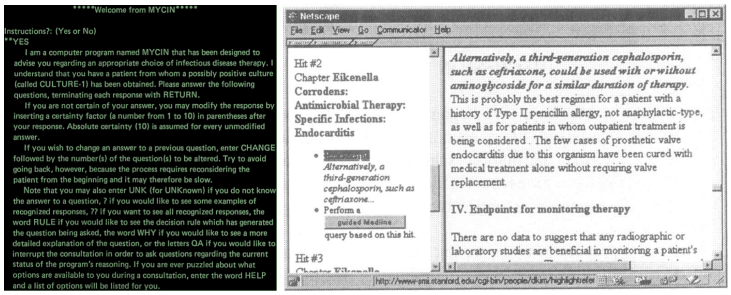
– les connaissances explicitées : éléments tangibles qui peuvent être formalisés (livres, écrits, mails, etc.),

– les connaissances tacites : éléments intangibles portés par les individus et qui ne peuvent pas toujours être explicités puisque « Nous savons plus que nous ne pouvons dire » (Polanyi, 1967).

Ainsi, tout comme les connaissances tacites et explicitées jouent un rôle crucial dans la transmission du savoir humain, l'IA cherche à reproduire et à simuler ces processus cognitifs complexes.

3.2. L'intelligence artificielle, des processus cognitifs simulés

MYCIN est un des premiers systèmes experts, c'est-à-dire logiciel utilisant une base de connaissances et un moteur d'inférence permettant de simuler des raisonnements humains : l' « intelligence » est « artificielle». Ce système était conçu pour extraire les connaissances d'un médecin, les intégrer dans un programme informatique et fournir des diagnostics, bien que non différentiels, à l'aide d'un arbre de décision (Shortliffe, 1976). La figure 4 présente l'interface de MYCIN en 1976 (a) ainsi qu'une tentative de MYCIN II en 1998 intégrant un langage de balisage (b).



(a) (Shortliffe, 1976) (b) (Kim *et al*., 1998)

FIGURE 4. *Les interfaces utilisateur de MYCIN en 1976 (a) et MYCIN II en 1998 (b)*

Ce qui peut être considéré comme la conférence fondatrice de l'intelligence artificielle est le workshop organisé en 1955 par McCarthy *et al*. (1955). L'idée était alors de décrire les « caractéristiques de l'intelligence de manière si précise qu'une machine puisse être conçue pour la simuler ». Ce document contient déjà beaucoup d'éléments sur le traitement du langage, l'abstraction et la créativité. Benchimol *et al*. (1986) proposent de considérer les systèmes experts comme des « programmes informatiques capables de reproduire des raisonnements humains », définition qui fait pleinement sens aujourd'hui pour mobiliser le concept d'IA, en particulier lorsqu'il est question de gestion des connaissances tacites. Cette notion de raisonnement revient également dans la définition données bien plus récemment par le Parlement Européen : « l'IA désigne la possibilité pour une machine de reproduire des comportements liés aux humains, tels que le raisonnement, la planification et la créativité. » [3]. L'IA se retrouve ainsi reliée de manière intrinsèque à la notion de connaissances, base même du raisonnement humain.

4. L'atelier « KM-IA »

La gestion des connaissances est un élément clé au sein des organisations (Di Vaio *et al*., 2021; Irani *et al*., 2009; Antunes, Pinheiro, 2020; Osman *et al*., 2022). Les

3. https://www.europarl.europa.eu/topics/fr/article/20200827STO85804/intelligence-artificielle-definition-et-utilisation

connaissances, une fois explicitées, contribuent à la mémoire organisationnelle et peuvent être considérées comme des actifs à part entière de l'organisation. Reste cependant la difficile question des connaissances tacites, dont le caractère intangible et souvent intuitif, rend complexe la transmission.

Dans ce contexte, se pose la question de l'apport que l'IA peut avoir en gestion des connaissances, notamment tacites. De manière générale, on peut légitimement se demander comment l'IA peut aider les entreprises à partager et pérenniser les connaissances tacites portées par les individus au sein des organisations.

Ce questionnement a été le point de départ pour l'atelier « KM-IA » qui s'est tenu début 2025 pendant la conférence EGC [4]. Celui-ci a réuni une quinzaine de participants, aussi bien issus du monde académique qu'industriel, autour de retours d'expériences et de discussions sur les apports et les limites de l'IA pour la gestion des connaissances tacites.

Lors de cet atelier, un total de 8 contributions ont été présentées. Celles-ci ont couvert différents domaines, comme la gestion de fraudes dans le domaine bancaire (Chergui *et al*., 2025), le domaine hospitalier (Tounkara, Ducert, 2025) et les géosciences (Dechambenoit *et al*., 2025), aussi bien dans l'industrie (Berger, Prieur, 2025), que dans le domaine public (Steffenel, Lucas, 2025).

Ainsi, Ben Khaled, Monticolo (2025) ont soulevé la question de la coopération IA-humain, en explorant les modèles de langages à grande échelle (LLM) pour l'interaction avec des experts et la formalisation des savoirs, dans le cadre de l'industrie 4.0. A partir d'une architecture microservices, le travail propose l'usage des modèles d'IA pour poser des questions aux experts et enrichir le contexte des connaissances en cours de formalisation. On retrouve ainsi une inversion des rôles, où l'IA pose les questions aux experts au lieu de répondre à leurs questions. L'IA s'affiche alors comme une aide au difficile travail d'explicitation des connaissances.

Ensuite, Rosenthal-Sabroux *et al*. (2025) ont considéré le rôle que l'IA pourrait avoir dans le transfert de connaissances tacites. Pour ces auteurs, l'IA n'a pas de connaissances à proprement parler, car celles-ci ne sont pas des objets, elles reposent en grand partie sur une expérience contextuelle et subjective difficile à formaliser par des algorithmes qui constituent la base de l'IA. Cependant l'IA pourrait contribuer au transfert des connaissances en aidant l'humain dans la difficile tâche d'expliciter des connaissances. Pour Rosenthal-Sabroux *et al*. (2025), "*la connaissance est une rencontre d'un sujet avec une donnée*". En effet, en entreprise, la connaissance va forcément être reliée à l'action, et notamment à travers le processus métiers au cours desquels elle est utilisée. Ainsi, la transmission des connaissances tacites se fait principalement par l'interaction avec quelqu'un d'autre : c'est lors de cette interaction que la connaissance émerge et pourra ainsi être explicitée et transmise. Cependant, encore faut-il que dans cette explicitation, la personne qui reçoit la connaissance puisse l'interpréter de la même manière que celle qui la transmet.

4. https://km-ia.sciencesconf.org/

La transmission des connaissances est également au centre des travaux de Tounkara, Ducert (2025). Ces auteurs ont présenté une étude de cas sur la gestion des connaissances et l'IA en milieu hospitalier. Cet étude part d'un projet pilote dont l'objectif est de faire évoluer le système de gestion de connaissances d'un CHU en y intégrant des outils d'IA afin d'améliorer l'efficacité de ses processus de gestion des connaissances (transfert, création et application de connaissances). L'étude a pu ainsi mettre en lumière des enjeux et des défis liés notamment à l'utilisation de l'IA au sein d'une structure hospitalière. Pour ces auteurs, l'IA est une réalité dans les hôpitaux. Cependant, il convient de séparer l'IA pour la santé (c.a.d. pour les usages médicaux) de celle pour les aspects administratifs, comme, par exemple la gestion des plannings et des flux. Dans le cadre de cet étude, ces sont ces aspects administratifs, essentiels pour le quotidien du personnel hospitalier (infirmiers, médecins et autres), qui étaient visés. Parmi les défis observés, nous pouvons citer la question du support aux processus organisationnels au sein d'un CHU et le manque d'outil d'aide à l'explicitation des connaissances. Au-delà de l'aspect réglementaire (RGPD, AI Act, etc.) nécessaire à la confiance, à la transparence et à la crédibilité des outils de gestion de connaissances comportant de l'IA dans ce milieu, les auteurs ont souligné l'importance de l'impact sociétal et la nécessité d'un positionnement sociologique de ce type de projet.

Le domaine financier est également concerné par ces réflexions. En effet, le nombre de transactions financières augmentant de manière significative, il devient de plus en plus difficile d'identifier des patrons de fraude. Dans ce contexte, l'IA peut s'avérer être un outil indispensable d'aide à la décision, soutenant le travail des experts. Dans son article, Chergui *et al*. (2025) proposent une approche hybride, mélangeant des techniques d'apprentissage automatique et des ontologies de domaines afin d'aider des experts dans l'interprétation des transactions suspectes et la définition, avec les experts, de règles caractérisant des schémas de fraude. Les techniques d'apprentissage automatique (*Machine Learning*) permettent l'analyse d'un grand volume des données, qui n'aurait pas pu être traité manuellement par les experts, et l'identification d'anomalies, qui pourraient constituer des nouveaux schémas de fraude. Cependant, il reste la question de l'interprétation des transactions identifiées comme schéma de fraude. La question est abordé ici en combinant, les ontologies aux techniques d'explicabilité comme SHAP. Les ontologies vont permettre de structurer les connaissances explicitées par les experts, facilitant leur transfert au sein des organisations, alors que les techniques d'explicabilité sont là pour renforcer la transparence, afin que les experts puissent comprendre et justifier les décisions prises.

A travers de projets comme ceux de Tounkara, Ducert (2025); Chergui *et al*. (2025); Dechambenoit *et al*. (2025), on observe l'impact potentiel que l'IA peut avoir sur les pratiques professionnels. Il se pose alors la question du rôle des experts métiers dans les projets IA instigateurs de nouvelles pratiques. Cette question est au centre de la contribution de Nesvijevskaia (2025), qui a analysé un corpus de sept projets impliquant la conception d'outils IA. Cette analyse a permis à son auteur de dégager plusieurs pistes de réflexion, dont l'importance du savoir tacite de l'expert métier pour définir, prioriser et juger les critères d'évaluation des modèles créés lors de ces projets. L'usage de l'IA dans l'explicitation des connaissances suscite également la question

du rôle d'intermédiaire entre l'expert et le réel que les modèles IA sont invités à jouer : est-ce que la présence de cet intermédiaire ne pourrait pas à terme nuire au développement de savoirs métier au profit de compétences liées à la manipulation des outils IA ? La question reste ouverte...

Comme Ben Khaled, Monticolo (2025), les travaux de Steffenel, Lucas (2025); Dechambenoit *et al*. (2025) cherchent dans l'IA une aide à l'explicitation des connaissances et à la pérennisation des savoirs autour des processus métiers. Steffenel, Lucas (2025) présente comment le projet de fédération ILaaS (Inférence LLM as a Service) prétend contribuer à cette question à travers la mutualisation des ressources et l'usage des RAG (*Retrieval-Augmented Generation*). En effet, les universités sont des organismes complexes, où les connaissances sont rarement bien codifiées et structurées. Les processus sont nombreux, comportent des nombreuses exceptions et sont souvent mal documentés. La transmission des connaissances liés à ces processus est, par conséquent, complexe, ce qui peut conduire à des problèmes de continuité de service et d'*onboarding*. Par ailleurs, les processus manipulent souvent de l'information à caractère sensible, allant des ressources humaines aux données de recherche. Dans ce contexte, l'usage des LLMs hébergées par des tiers peut poser problème, tout comme la question du contrôle d'accès à ces informations. Les auteurs soutiennent donc l'usage des RAG au sein d'une fédération mutualisant des ressources de calcul, ce qui permettrait de garder les informations localement dans la fédération et de gérer les niveaux d'accès aux informations.

Dechambenoit *et al*. (2025), quant à eux, s'intéressent à la transformation des savoirs individuels, souvent implicites et non formalisés, en un patrimoine collectif, structuré et exploitable, au sein des organisations scientifiques, comme, par exemple, le Bureau de Recherches Géologiques et Minières (BRGM). L'IA est utilisée ici pour recenser et retranscrire ces savoirs sous forme procédurale, en analysant les notes de terrain, les échanges informels et les pratiques géoscientifiques au BRGM. A nouveau, le caractère personnel et difficilement accessible des savoirs tacites est au centre des discussions. On retrouve ainsi la place de l'intuition dans la pratique des experts sur le terrain et l'importance de l'action pour la connaissance, déjà mise en avant par Rosenthal-Sabroux *et al*. (2025). Aussi la variabilité entre les pratiques des experts représente ici un défis supplémentaire dans la formalisation de pratiques sous la forme de processus.

Enfin, la contribution de Berger, Prieur (2025) s'est intéressée à la mémoire d'entreprise et aux solutions type "base de connaissances". Pour ces auteurs, la mise en place d'un Système de Management de la Connaissance est un long chemin ardu, dont les résultats peuvent être particulièrement positifs pour les organisations, comme le renforcement de l'identité culturelle de l'organisation ou encore l'amélioration des échanges par l'établissement d'un langage commun. Dans ce cadre, l'usage des LLM pour l'interrogation de grandes bases de connaissances apparaît comme une possibilité intéressante, mais qui est confrontée à la question de l'accès à l'information. En effet, ces bases de connaissances contiennent la mémoire stratégique de l'entreprise, et peuvent difficilement être connectées à l'extérieur de celle-ci. L'usage de RAG, tel

que soutenue aussi par Steffenel, Lucas (2025), est à nouveau mise en avant. Pour ces auteurs, la "Mémoire d'Entreprise" est un sujet stratégique, qui doit se construire collectivement, à travers aussi bien des techniques plus "traditionnelles" de l'Ingénierie des Connaissances que des nouvelles technologies comme les RAG et les LLM.

5. Les apports et limites de l'IA pour gérer les connaissances tacites

L'appel à contributions de l'atelier évoquait des questions comme :

– Comment l'IA peut aider les entreprises à partager et pérenniser les connaissances tacites portées par les individus au sein des organisations ?

– Quels usages des outils basés sur l'IA pour analyser les interactions et les communications afin d'en extraire des connaissances (analyse textuelle des comptes-rendus des réunions, des notes de réunions, des enregistrements vidéo de réunions, des données issues d'outils collaboratifs de l'entreprise, etc.) ?

– Vers quelles directions pointent les premiers retours d'expérience en entreprise ? Quelles sont les implications de l'application de l'IA sur l'organisation et ses employés ?

– Quelles limites managériales et éthiques émergent ?

En effet, s'il convient de s'intéresser aux apports de l'IA pour les organisations et la société, il convient surtout de ne pas négliger ses limites (Kilovaty, 2025). Ces deux dimensions, d'apports et de limites, vont être abordées dans la suite de cette section.

5.1. Sur les apports de l'IA pour la gestion des connaissances tacites

L'ensemble des participants a soulevé des apports indéniables de l'IA pour la gestion des connaissances tacites.

5.1.1. Capture des mécanismes de raisonnement

La capture des mécanismes de raisonnement est par exemple abordée par Ben Khaled, Monticolo (2025) comme un moyen de modélisation et de traçabilité des processus décisionnels qui permet par ailleurs une organisation hiérarchique des savoirs, y compris des connaissances tacites au travers de l'adaptation aux profils d'apprentissage et aux schémas mentaux. En effet, au travers de l'itération de prompts successifs, l'utilisateur est guidé pour clarifier et structurer sa pensée, transformant ainsi sa connaissance tacite en connaissance explicite (Rosenthal-Sabroux *et al*., 2025). Pour Nesvijevskaia (2025), la formalisation de la problématique métier est rendue possible par la capture des connaissances tacites issues de la pratique pour les intégrer à l'usage. Les savoirs métier sont traduits en modèles data, révélant parfois des biais cognitifs. L'impact de la capture peut se révéler majeur par la diminution puis l'élimination du « risque de personne-clé », tout en pouvant révéler parfois de possibles incohérences. Pour Dechambenoit *et al*. (2025), il y a une fenêtre précise pour le « prélèvement du tacite ».

5.1.2. Amélioration de la documentation

La réduction des temps de capture permet une amélioration de la qualité documentaire qui entraîne l'optimisation des processus de formation pour Ben Khaled, Monticolo (2025). L'IA peut en effet enrichir et améliorer sa base de connaissances qu'elle sera en mesure de mobiliser dans des interactions futures (Rosenthal-Sabroux *et al.*, 2025) : l'utilisation collective de l'IA pour accéder à la connaissance peut aboutir à la création d'un ensemble de connaissances tacites partagées, formant ainsi une culture commune. Tounkara, Ducert (2025) évoquent le contexte hospitalier où la mise à jour automatique des documents à partir des connaissances explicitées a été réalisée avec de l'IA via un processus guidé de questions à des médecins experts pour expliciter les connaissances tacites, alors que Chergui *et al*. (2025) présentent la construction d'une ontologie comme un moyen de hierarchiser les concepts et de capturer les connaissances tacites (Fig. 5).

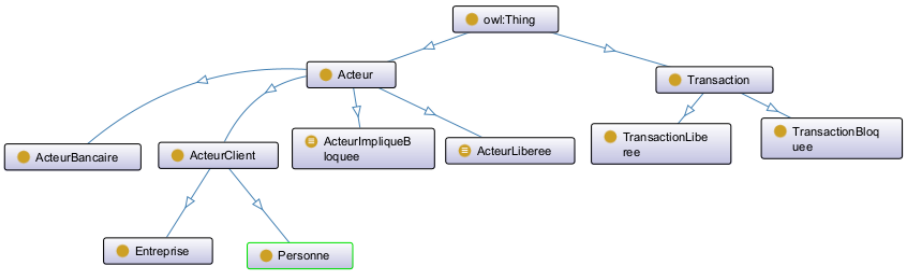


FIGURE 5. *Capture de connaissances basée sur une ontologie (Chergui* et al.*, 2025)*

5.1.3. Création de pratiques nouvelles et innovation

Intégrer un nouvel artefact peut avoir un effet catalyseur sur la création de connaissances et de pratiques nouvelles pour Dechambenoit *et al*. (2025). Berger, Prieur (2025) vont plus loin en présentant un portail en ligne permettant de déployer un système de management de la connaissance en entreprise qui vise à pérenniser, expliciter, transmettre et exploiter les connaissances selon la norme ISO 30401:2018.

5.2. Sur les limites de l'IA pour la gestion des connaissances tacites

Bien qu'évoquées par des auteurs comme Mikalef *et al*. (2022), la part sombre et les mauvaises surprises de l'IA restent sous-étudiées. Cela a d'ailleurs été en partie le point de départ de l'atelier « KM-IA ».

5.2.1. Risque d'apprendre des erreurs

L'explicitation reste parfois superficielle et les connaissances peuvent ne pas être saisies dans toute leur complexité ou spécificité Tounkara, Ducert (2025). Certains aspects de la connaissance, comme l'historique avec les justificatifs d'évolution, ne

sont par ailleurs pas toujours captés. L'intelligence artificielle (IA) joue un rôle crucial dans l'interaction avec les experts et la complétion des bases de connaissances. Cependant, il existe un risque de lacunes, car l'explicitation des connaissances tacites repose fortement sur l'interaction humaine.

Le problème de maintenir les connaissances à jour se pose. Pour Dechambenoit *et al*. (2025), cela est possible avec le versionnage, dans le cadre du versionnage des processus. Selon eux, cette approche permet de poursuivre une intuition pour réaliser une tâche différemment, d'intégrer un nouvel outil qui fait évoluer la pratique, de tenter d'optimiser le processus existant, ou encore de s'adapter à un contexte spécifique. Ces différentes versions représentent diverses approches pour atteindre un même objectif, chacune enrichissant le corpus des pratiques possibles.

5.2.2. Hébergement par des tiers et aspects réglementaires

Considérant les travaux de Mollick (2024), Steffenel, Lucas (2025) pointent du doigt la nécessaire prise de conscience sur le risque d'un hébergement par des tiers. La formation de services souverains, en particulier sous forme de fédération, apparaît alors comme une solution également présentée comme un moyen d'articulation entre tous les professionnels de santé sur un territoire par Tounkara, Ducert (2025). Les réglementations proposées par l'Union Européenne vont dans le sens d'une prise en compte de cette nécessité (Rosenthal-Sabroux *et al*., 2025), bien qu'un cadre réglementaire devrait également poser des questions sur les données des salariées en entreprise.

5.2.3. Biais et acceptabilité des solutions

Tounkara, Ducert (2025) défendent que la mobilisation de formes tacites et collectives des connaissances peut permettre de réduire les biais liés à la nature incertaine de l'environnement, à la crédibilité des données/informations, à leur caractère incomplet et à l'éthique des algorithmes. Néanmoins, des doutes persistent sur les mécanismes organisationnels permettant de mobiliser les réseaux formels et informels afin de vérifier les informations sources de connaissances émanant de l'IA. Steffenel, Lucas (2025); Chergui *et al*. (2025) insistent sur l'acceptabilité organisationnelle des solutions qui requièrent en effet des mises à jour régulières à des fins de vérification par un agent humain. Si la pluridisciplinarité est essentielle pour faire cohabiter les impératifs métier avec les besoins de réflexion nécessaires à l'adoption de l'IA, il est également crucial de réfléchir aux impacts sociétaux de cette adoption, afin de garantir une intégration harmonieuse et bénéfique pour tous.

En effet, les difficultés liées à l'explicabilité des propositions faites par l'IA sont nombreuses, notamment en ce qui concerne la construction et la mise à jour des ontologies (Chergui *et al*., 2025). L'IA, et plus spécifiquement le *Machine Learning* (ML), doivent être vus comme des outils d'aide à la décision, mais la décision finale doit toujours revenir à l'humain. Rosenthal-Sabroux *et al*. (2025) rappellent que « *la connaissance n'est pas un objet* » ; elle peut être tacite et peut alors être transmise dans l'interaction, par compagnonnage par exemple. L'IA peut aider à expliciter ces

connaissances tacites, mais elle ne peut pas comprendre les conditions de leur émergence. Les influences culturelles, l'expérience et les croyances jouent un rôle crucial dans la formation de ces connaissances, tout comme l'intuition. Il est donc important de reconnaître les limites de l'IA dans ce domaine et de valoriser l'apport humain dans la gestion des connaissances tacites.

5.3. Discussion

Il convient de noter à présent que l'absence de « bon sens » ou d' « intuition » n'ont pas été rapportées comme une limite de l'IA pour gérer les connaissances tacites. De même, le manque de compréhension du contexte est crucial, l'IA pouvant avoir du mal à comprendre le contexte et les nuances des connaissances tacites, basées sur la pratique et l'expérience personnelle.

(Polanyi, 1969, p. 195) va plus loin lorsqu'il écrit que : « L'idée d'une connaissance strictement explicite est [...] contradictoire ; privées de leurs coefficients tacites, toutes les paroles, toutes les formules, toutes les cartes et tous les graphiques sont strictement dénués de sens. ». Se pose alors la question de savoir non pas s'il faut utiliser l'IA pour formaliser les connaissances tacites, ce qui les prive de leur sens, mais plutôt de savoir quel serait le bon usage de l'IA pour participer à la gestion des connaissances tacites ?

L'IA peut être un outil puissant d'interaction avec non seulement des experts, porteurs de connaissances tacites cruciales pour les entreprises, mais aussi avec des néophytes dans une perspective d'apprentissage par l'échange. Sans aller jusqu'à devenir un mentor, l'IA peut participer à la gestion des connaissances tacites par son caractère interactionnel fort, poussant à la formulation et à la reformulation de prompts successifs qui amènent l'utilisateur à clarifier et structurer sa pensée.

Une autre limite à prendre en compte est le risque d'uniformisation des connaissances tacites dues à une utilisation collective de l'IA. La construction d'un ensemble de connaissances tacites partagées par les utilisateurs peut former une culture commune et universelle qui, alors qu'elle servira de base de connaissances pour les échanges futurs, risque d'amoindrir le potentiel de création de connaissances disruptives.

Néanmoins, les années d'efforts consacrés à l'ingénierie des connaissances ne sont pas à rejeter. Cette discipline a proposé des modèles structurants pour faciliter l'interaction entre les experts et les ingénieurs en connaissances ; ils servaient de guide précieux pour ces échanges, les ontologies en étant un exemple notable. Par son potentiel d'intégration des connaissances et sa capacité interactionnelle, l'ingénierie des connaissances apparaît alors, sinon comme une solution, comme un moyen de prendre en compte les connaissances tacites dans leur sphère la plus humaine : l'interaction.

6. Conclusions et perspectives

L'explicitation des connaissances tacites ne suffit pas à elle seule pour les gérer. L'IA peut potentiellement jouer un rôle crucial dans la compréhension de ces connais-

sances tacites, en aidant à les interpréter et à les contextualiser. Cependant, il est important de noter que certains experts peuvent être réticents à s'investir dans ce processus d'explicitation, craignant que cela ne mette en péril leur emploi. Cette résistance au changement souligne la nécessité d'une gestion attentive de la prise en compte de l'humain dans les processus des gestion des connaissances tacites en entreprise.

Enfin, la pluridisciplinarité est essentielle pour une réflexion approfondie sur les impacts sociétaux de l'IA. Dans les grandes organisations, notamment publiques, la question de la pérennité des ressources nécessaires à la maintenance des solutions mises en place est cruciale. L'IA peut également aider à fournir des interprétations des connaissances en fonction de leur usage, reconnaissant que le sens peut varier selon les individus et leurs propres expériences. Cette capacité à adapter les interprétations en fonction des contextes d'utilisation renforce l'importance de l'IA comme outil d'aide à la décision, la décision revenant toujours et devant toujours au final revenir à l'humain.

Bibliographie

Antunes H. d. J. G., Pinheiro P. G. (2020). Linking knowledge management, organizational learning and memory. *Journal of Innovation & Knowledge*, vol. 5, n° 2, p. 140-149. Consulté sur https://www.sciencedirect.com/science/article/pii/S2444569X19300319

Arduin P.-E. (2018). *La menace intérieure* (vol. 9). ISTE Group.

Arduin P.-E., Ziam S. (2024). If digital tools are the solution to knowledge transfer, what is the problem? In S. P. Duarte, A. Lobo, B. Delibašić, D. Kamissoko (Eds.), *Decision support systems xiv. human-centric group decision, negotiation and decision support systems for societal transitions*, p. 126–138. Cham, Springer Nature Switzerland.

Benchimol G., Lévine P., Pomerol J.-C. (1986). *Systèmes experts dans l'entreprise*. Hermes.

Ben Khaled K., Monticolo D. (2025). Vers une coopération ia-humain pour la capture et la transmission des savoir-faire métier. In *Atelier "gestion des connaissances tacites en entreprise : réflexions, retours d'expériences, bonnes pratiques et mauvaises surprises de l'intelligence artificielle", conférence egc 2025, strasbourg, 28 janvier.*

Berger A., Prieur P. (2025). Le « management de la connaissance » : la clé stratégique de la réflexion sur l'apport de la « mémoire d'entreprise ». In *Atelier "gestion des connaissances tacites en entreprise : réflexions, retours d'expériences, bonnes pratiques et mauvaises surprises de l'intelligence artificielle", conférence egc 2025, strasbourg, 28 janvier.*

Chergui H., Abrouk L., Cabioch N. (2025). L'intelligence artificielle pour la gestion des connaissances et la lutte contre la fraude financière dans les institutions internationales. In *Atelier "gestion des connaissances tacites en entreprise : réflexions, retours d'expériences, bonnes pratiques et mauvaises surprises de l'intelligence artificielle", conférence egc 2025, strasbourg, 28 janvier.*

Dechambenoit G., Chamekh F., Laouici I., Dantal Y., Loiselet C. (2025). les géosciences face au challenge des savoirs tacites : retour d'experience et perspectives. In *Atelier "gestion des connaissances tacites en entreprise : réflexions, retours d'expériences, bonnes pratiques et mauvaises surprises de l'intelligence artificielle", conférence egc 2025, strasbourg, 28 janvier.*

Di Vaio A., Palladino R., Pezzi A., Kalisz D. E. (2021, février). The role of digital innovation in knowledge management systems: A systematic literature review. *Journal of Business Research*, vol. 123, p. 220–231.

Irani Z., Sharif A. M., Love P. E. (2009). Mapping knowledge management and organizational learning in support of organizational memory. *International Journal of Production Economics*, vol. 122, n° 1, p. 200-215. (Transport Logistics and Physical Distribution Interlocking of Information Systems for International Supply and Demand Chains Management ICPR19)

Kilovaty I. (2025). Hacking generative ai. *Loyola of Los Angeles Law Review*, vol. 58.

Kim D. K., Fagan L. M., Jones K. T., Berrios D. C., Yu V. L. (1998). Mycin ii: design and implementation of a therapy reference with complex content-based indexing. In *Proceedings of the amia symposium*, p. 175.

Kirsch-Pinheiro M. (2023). The context awareness challenges for pis. In *The evolution of pervasive information systems*, p. 43–63. Springer.

McCarthy J., Minsky M. L., Rochester N., Shannon C. E. (1955). A proposal for the dartmouth summer research project a proposal for the dartmouth summer research project on artificial intelligence. *Project Proposal*.

Mecharnia T., Khelifa L. C., Hamdi F., Pernelle N., Rouveirol C. (2021). Découverte de règles contextuelles pour prédire la présence d'amiante dans les bâtiments. In *Journées francophones d'ingénierie des connaissances (ic) plate-forme intelligence artificielle (pfia'21)*, p. pp–73.

Mikalef P., Conboy K., Lundström J. E., Popovič A. (2022). *Thinking responsibly about responsible ai and 'the dark side' of ai* (vol. 31) n° 3. Taylor & Francis.

Mollick E. (2024). *Co-intelligence: Living and working with ai*. Penguin Publishing Group.

Nesvijevskaia A. (2025). Pérenniser le savoir tacite des experts métier à travers les projets d'ia : retours d'expérience. In *Atelier "gestion des connaissances tacites en entreprise : réflexions, retours d'expériences, bonnes pratiques et mauvaises surprises de l'intelligence artificielle", conférence egc 2025, strasbourg, 28 janvier.*

Nonaka I., Takeuchi H. (1995). *The knowledge-creating company*. Oxford University Press.

Osman M. A., Noah S. A. M., Saad S. (2022). Ontology-based knowledge management tools for knowledge sharing in organization—a review. *IEEE Access*, vol. 10, p. 43267-83.

Polanyi M. (1967). Sense-giving and sense-reading. *Philosophy: Journal of the Royal Institute of Philosophy*, vol. 42, n° 162, p. 301-323.

Polanyi M. (1969). *Knowing and being*. London, Routledge and Kegan Paul.

Rosenthal-Sabroux C., Negre E., Mayag B., Jaillet T. (2025). L'intelligence artificielle numérique face au défi des connaissances tacites humaines. In *Atelier "gestion des connaissances tacites en entreprise : réflexions, retours d'expériences, bonnes pratiques et mauvaises surprises de l'intelligence artificielle", conférence egc 2025, strasbourg, 28 janvier.*

Shortliffe E. (1976). Books: Computer-based medical consultations: Mycin. *Journal of Clinical Engineering*, vol. 1, n° 1, p. 69.

Steffenel L.-Z., Lucas L. (2025). L'intérêt des rag dans la gestion des connaissances des processus administratifs universitaires à l'ère des llm. In *Atelier "gestion des connaissances tacites en entreprise : réflexions, retours d'expériences, bonnes pratiques et mauvaises surprises de l'intelligence artificielle", conférence egc 2025, strasbourg, 28 janvier.*

Tounkara T., Ducert D. (2025). Systèmes de gestion des connaissances et intelligence artificielle dans le contexte hospitalier : enjeux et défis. In *Atelier "gestion des connaissances tacites en entreprise : réflexions, retours d'expériences, bonnes pratiques et mauvaises surprises de l'intelligence artificielle", conférence egc 2025, strasbourg, 28 janvier.*

OSDN, une plateforme pour la recherche interdisciplinaire en Science Ouverte

Vincent-Nam Dang [1], Nathalie Aussenac-Gilles [2], Imen Megdiche[3] Franck Ravat [1]

1. IRIT, CNRS (UMR 5505), Université Toulouse Capitole, France
Vincent-nam.Dang@irit.fr, Franck.Ravat@irit.fr

2. IRIT, CNRS (UMR 5505), France
Nathalie.aussenac-gilles@irit.fr

3. IRIT, CNRS (UMR 5505), INU Champollion, ISIS Castres, Université de Toulouse, France
Imen.Megdiche@irit.fr

REFERENCE DE L'ARTICLE INTERNATIONAL Vincent-Nam Dang, Nathalie Aussenac-Gilles, Imen Megdiche, Franck Ravat. Enabling Interdisciplinary Research in Open Science: Open Science Data Network. In: Araújo, J., de la Vara, J.L., Santos, M.Y., Assar, S. (eds) Research Challenges in Information Science. RCIS 2024. LNBIP, vol 513. Springer. 19-34.

1. Introduction

La science ouverte est un mouvement qui vise à améliorer le processus de création de connaissances pour permettre la collaboration entre différentes communautés de recherche, qu'elles soient inter- ou intra-disciplinaires. La science ouverte est adossée aux principes FAIR, qui décrivent quatre piliers pour améliorer le partage et l'accès aux données. Parmi ces principes, la "trouvabilité" est définie comme la facilité à trouver des données pour les humains et les machines (Jacobsen A. et al, 2020). Les chercheurs de différents domaines et communautés soulignent d'importants freins lorsqu'ils recherchent des données brutes ou des jeux de données résultant de travaux de recherche. Ces freins sont essentiellement liés aux : (i) modèles de métadonnées des sources de données peu interopérables, (ii) manque de coordination entre les initiatives disciplinaires conduisant à un nombre très important de plateformes de gestion des données de recherche. L'un des moyens de résoudre ces problèmes consiste à mettre en place une plateforme centralisée de gestion des données pour la Science Ouverte. Mais la centralisation pose plusieurs problèmes : (i) impossibilité de répondre aux besoins spécifiques des chercheurs des

différentes communautés ; (ii) volume trop important pour une seule plateforme ; (iii) coût de déploiement trop élevé d'une plateforme qui deviendrait un point de fragilité ; (iv) aucune garantie de sécurité et de confidentialité.

2. La plateforme OSDN

La plateforme Open Science Data Network (OSDN) que nous proposons consiste en un réseau décentralisé, fédéré et distribué de plateformes de gestion de données de la Science Ouverte. Cette solution est basée sur une API RESTful utilisant un registre partagé par toutes les plateformes, qui contient des informations relatives aux plateformes (nom, URL, plateformes interconnectées, etc.), aux modèles de métadonnées utilisés (nom, contenu, etc.) et aux correspondances entre les modèles de métadonnées. Un mécanisme de propagation des modifications les diffuse à tous les voisins jusqu'à atteindre l'ensemble du réseau. Ce module prend la forme d'un conteneur Docker à déploiement automatique dont l'intégration ne nécessite que l'installation de la fonction d'interopération entre ce module et le mécanisme de recherche d'information de la plateforme. Choisir la bonne topologie du réseau est déterminant pour bien gérer les suppressions volontaires et involontaires. Une solution qui minimise la vulnérabilité à ces deux événements est la topologie de réseau sans échelle avec un seul hub, les autres nœuds ayant tous un degré de cinq. Pour évaluer la faisabilité de l'OSDN, nous avons développé une preuve de concept qui intègre 3 plateformes de Science ouverte avec différentes technologies de gestion des métadonnées. Nous avons exécuté une requête sur une plateforme et vérifié la propagation de cette requête aux plateformes voisines, puis aux voisins indirects. Pour garantir la gestion des modèles de métadonnées, nous avons intégré et mis en relation 19 modèles de métadonnées dans le registre de l'OSDN. Nous avons créé manuellement des correspondances entre ces modèles structurels afin d'assurer leur interopérabilité. Pour l'expérimentation de OSDN, un chercheur spécialiste en agronomie a exploité la plateforme pour chercher des jeux de données décrivant une activité de biocontrôle destinée aux agriculteurs. Plus précisément, il cherchait des données relatives à la la souche « trichoderma harzianum T-22 ». Grâce à l'utilisation de l'OSDN, le temps de recherche est réduit de 80 % ; il accède à un plus grand nombre de jeux de données (+ 7%). Dans le cadre des travaux futurs, nous envisageons de réduire le coût d'adoption de cette solution et de mieux exploiter l'interopérabilité sémantique. Enfin, nous voulons mieux assurer le passage à l'échelle, en particulier en optimisant la consommation de ressources dans l'OSDN.

3. Références

Jacobsen A. et al. (2020). Fair principles: interpretations and implementation considerations. Data intelligence, vol. 2, no 1-2, p. 10–29

Tendances de recherche sur la convergence des grands modèles de langage et des graphes de connaissance

Hanieh Khorashadizadeh[1], Fatima Zahra Amara[2] , Morteza Ezzabady[3], Frédéric Ieng[4], Sanju Tiwari[5], Nandana Mihindukulasooriya[6], Jinghua Groppe[1], Soror Sahri[4], Farah Benamara[3], Sven Groppe[1]

1. IFIS, University of Lübeck, Ratzeburger Allee 160/Haus 1, 23562, Lübeck, Germany
hanieh.khorashadizadeh@uni-luebeck.de, jinghua.groppe@uni-luebeck.de, sven.groppe@uni-luebeck.de

2. University of Bari, Via E. Orabona, 4 - 70125 Bari, Italy
fatima.amara@uniba.it

3. IRIT, Université de Toulouse, 118 Rte de Narbonne, 31400, Toulouse, France
morteza.ezzabady@irit.fr, farah.benamara@irit.fr

4. LIPADE, Université Paris Cité, 45 rue des Saints-Pères, 75006 Paris, France
frederic.ieng@u-paris.fr, soror.sahri@parisdescartes.fr

5. Alliance University, Bangalore, India
tiwarisanju18@ieee.org

6. IBM Research, 1101 Kitchawan Rd, Yorktown Heights, NY 10598, New York, US
nandana@ibm.com

Référence de l'article international. Cet article est une synthèse de l'article :
Research Trends for the Interplay between Large Language Models and Knowledge Graphs. VLDB 2024 Workshop: LLM+KG.

1. Introduction

Ces dernières années, les progrès rapides de l'intelligence artificielle ont été stimulés par deux technologies essentielles : les grands modèles de langage (LLM) et les graphes de connaissances (KG). Les graphes de connaissances fournissent une représentation structurée des connaissances, permettant aux machines de stocker, d'extraire et de déduire les relations sémantiques entre les entités. Parallèlement, les

LLM, tels que la série GPT de OpenAI, ont démontré des capacités remarquables en matière de traitement du langage naturel (NLP), de génération de contenu et de raisonnement.

Cependant, ces grands modèles souffrent souvent d'hallucinations, d'incohérences et d'un manque de représentation structurée des connaissances. Les graphes de connaissances, quant à eux, sont confrontés à des défis en matière d'évolution et d'adaptabilité. La combinaison de ces deux technologies pourrait permettre de créer des systèmes d'intelligence artificielle plus efficace, intelligents et faciles à interpréter.

2. Relations entre grands modèles de langue et graphes de connaissances

Cette étude explore la relation entre les LLM et les graphes de connaissances, en classant leurs interactions en trois domaines clés : LLM pour les graphes de connaissances, LLM améliorés par les graphes de connaissances et Coopération LLM-graphes de connaissances (Fig. 1).

Les LLM peuvent améliorer les graphes de connaissances de différentes manières : construction du graphe de connaissances, enrichissement d'un graphe de connaissances déjà existant, amélioration des raisonneurs, calcul de la représentation vectorielle du graphe de connaissances (KG embedding), etc.

De plus, les LLM peuvent aussi voir leurs performances améliorées par les graphes de connaissances avec l'ajout d'une source d'information externe. Elle permet ainsi d'améliorer les performances du modèle dans la compréhension de texte et de réduire le phénomène d'hallucination.

Enfin, les LLM et les graphes de connaissances peuvent être utilisés conjointement pour plusieurs tâches. En effet, afin de retrouver des informations sur les graphes de connaissances plus facilement, il est possible de générer des requêtes SPARQL/CYPHER à partir de texte en langage naturel. Il existe également des agents conversationnels améliorés qui utilisent les deux technologies pour fournir des réponses plus justes.

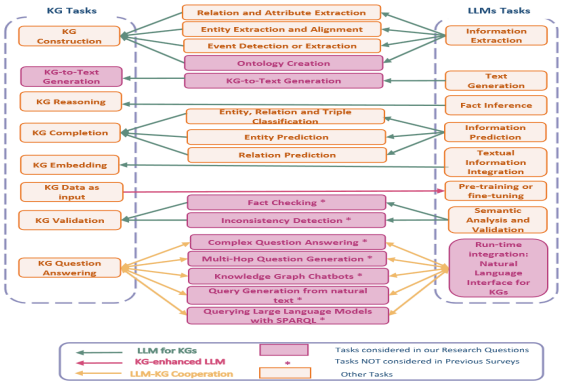


Figure 1. Les différentes synergies entre LLM et graphes de connaissances

De l'image à la représentation structurée : analyse et modélisation des manuels scolaires

Mohamed-Amine Lasheb [1], **Olivier Pons** [1],
Mohammed Bekkouche [2], **Isabelle Barbet** [1], **Caroline Huron** [3]

1. *Laboratoire Cedric, Conservatoire national des Arts et Métiers*
   292 Rue Saint-Martin, 75003 Paris, France
   firstname.lastname@lecnam.net
2. *LabRI-SBA, Ecole Supérieure en Informatique*
   22000 Sidi Bel Abbes, Algeria
   m.bekkouche@esi-sba.dz
3. *SEED, Inserm, Paris Cité*
   Paris, France
   firstname.lastname@cri-paris.org

RÉSUMÉ. *Le projet ANR MALIN vise à améliorer l'accessibilité des manuels scolaires numériques pour les élèves en situation de handicap en automatisant l'ensemble du processus d'adaptation. Une étape cruciale de ce processus est l'extraction et la structuration du contenu (leçons, illustrations, exercices, etc.). Cet article présente une solution basée sur des techniques de vision par ordinateur et d'apprentissage profond et compare l'efficacité de différents modèles.*

ABSTRACT. *The ANR MALIN project aims to improve the accessibility of digital textbooks for students with disabilities by automating the entire adaptation process. A crucial step in this process is the extraction and structuring of content (lessons, illustrations, exercises, , etc.). This paper presents a solution based on computer vision and deep learning techniques and compares the effectiveness of different models.*

MOTS-CLÉS : *éducation inclusive, accessibilité, extraction de contenu, structuration, modèles de vision par ordinateur.*

KEYWORDS: *inclusive education, accessibility, content extraction, structuring, computer vision models.*

1. Introduction

L'inclusion des enfants en situation de handicap dans les écoles et établissements scolaires ordinaires a été posée par la loi (lois du 11 février 2005 et du 8 juillet 2013) France (2005, 2013), ce qui a permis d'augmenter le nombre d'enfants en situation de handicap inscrits dans leur école de référence. Cependant, sa mise en place n'est pas simple. Un point d'achoppement concerne notamment les manuels scolaires, très utilisés en classe, mais qui, même dans leur version numérique (lorsqu'elles existent), sont très rarement accessibles.

Des adaptations sont généralement faites à la main par des associations et des organismes spécialisés, mais la diversité des manuels et leur renouvellement fréquent rendent ces adaptations lentes, coûteuses et peu nombreuses.

Sur la scène internationale, de nombreux pays ont introduit des obligations légales d'accessibilité minimale pour leurs livres scolaires. À l'échelle mondiale, c'est l'objet de l'initiative "Accessible Digital Textbooks", portée par l'UNICEF [1].

L'objectif de rendre accessibles les manuels scolaires en automatisant le processus de transposition, puis de permettre l'évaluation et l'amélioration des adaptations via la mise à disposition d'une plateforme d'adaptation, est donc un défi sociétal majeur.

L'objectif du projet ANR MALIN, dans lequel s'inscrit notre travail, est de répondre à ce défi. Du fait des collaborations avec l'association Le Cartable Fantastique [2] et l'INJA [3], le focus est principalement mis sur la dyspraxie et la déficience visuelle, mais le projet vise à se généraliser à tout type de handicap.

La Figure 1 montre des adaptations réalisées pour des élèves dyspraxiques. Elles visent à minimiser les tâches trop coûteuses pour eux, notamment l'écriture.

Une fois adaptés, les manuels peuvent être mis à disposition des publics en situation de handicap.

La structure complexe des manuels scolaires, illustrée dans la Figure 2, complique sensiblement leur adaptation. L'extraction et la structuration des contenus sont une première étape cruciale vers une automatisation des adaptations visant à rendre les manuels accessibles.

Ce travail explore l'utilisation de la vision par ordinateur pour automatiser l'extraction des contenus, permettant ainsi une représentation structurée des éléments des manuels.

1. https://www.accessibletextbooksforall.org/
2. https://www.cartablefantastique.fr/
3. Institut National des Jeunes Aveugles

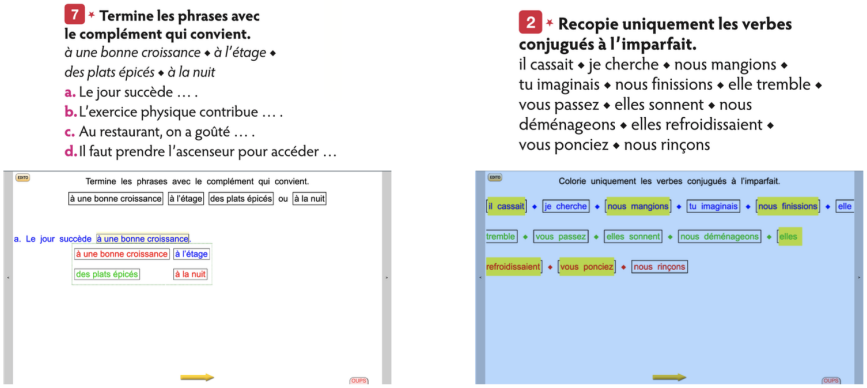


FIGURE 1 – Exemples d'adaptations d'exercices

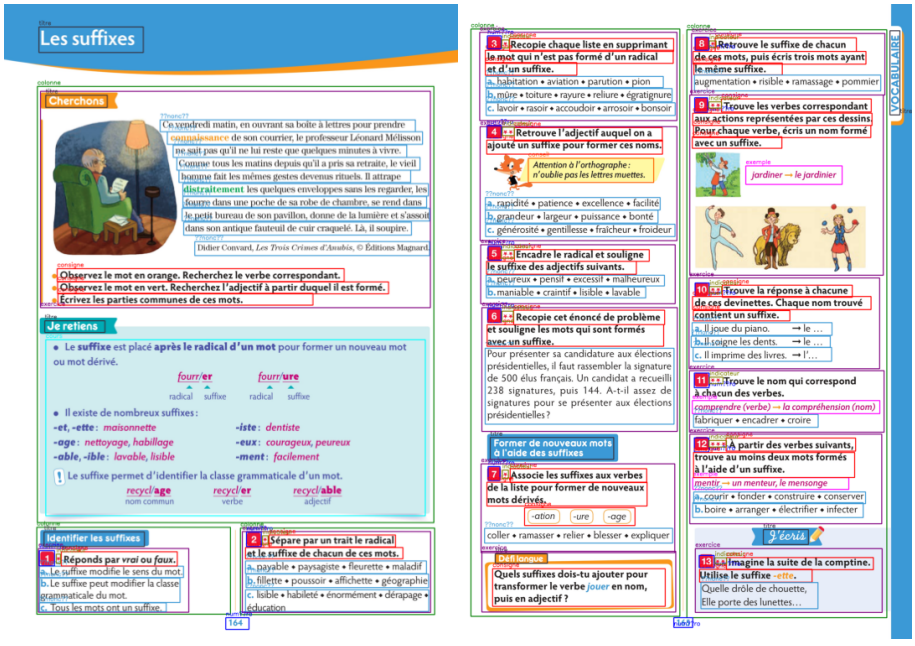


FIGURE 2 – Exemple de structure de mise en page d'un manuel scolaire. Source: Magnard (2019)

2. État de l'Art

Pour transformer les manuels scolaires en supports accessibles et interactifs, l'équipe du projet MALIN a proposé un pipeline complet, illustré dans la Figure 3. Ce pipeline décrit les étapes nécessaires à la conversion d'un fichier PDF, qu'il soit natif ou scanné,

en un manuel adapté au format HTML. On remarque que les deux dernières étapes du diagramme sont représentées de manière plus estompée : cela reflète le fait que ce travail se concentre principalement sur les deux premières, à savoir la collecte de données et la modélisation de la mise en page.

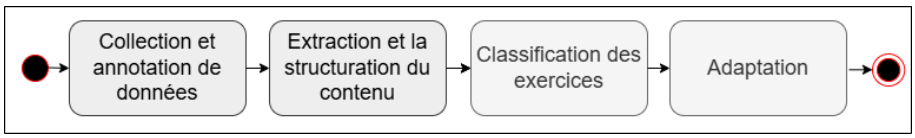


FIGURE 3 – Diagramme d'activité du processus global d'adaptation.

Ce processus débute par l'entrée d'un fichier PDF et aboutit à la création d'un manuel interactif. Les étapes clés incluent :

– **L'extraction et la structuration du contenu** des manuels scolaires. La sortie de cette étape est le point d'entrée de toutes les autres. Elle repose sur des modèles sémantiques formalisés dans Lincker, Pons *et al*. (2023). Ces formalismes s'appuient sur des DTD et des schémas XML ou JSON, et peuvent être traduits vers les normes TEI ((Rahtz *et al*., 2004; Stahn *et al*., 2016)) et DocBook (Walsh, Hamilton (2010)).

Après la constitution d'un premier corpus par des méthodes basées sur des règles et des approches statistiques, s'appuyant principalement sur les caractéristiques des polices et la position dans le document, un corpus de base a pu être établi, corrigé et étendu manuellement, bien qu'il ne puisse être diffusé publiquement en raison de restrictions liées aux droits d'auteur.

En utilisant ce corpus pour l'entraînement, des méthodes de TAL, utilisant des LLM basés sur des transformers et des transformers multimodaux (BERT Devlin *et al*. (2019), LayoutLM Xu *et al*. (2020), ViLa J. Lin *et al*. (2024)), ont été proposées dans Lincker, Pons *et al*. (2023).

– **La classification des exercices:** elle représente une autre étape essentielle. Les adaptations jouent un rôle clé dans ce processus. L'objectif est de classer chaque exercice en fonction du type d'adaptation le plus approprié. Pour ce faire, Lincker, Guinaudeau *et al*. (2023) s'appuient sur des modèles de langage pré-entraînés Martin *et al*. (2020); Le *et al*. (2020), et tire parti d'architectures multimodales Xu *et al*. (2022); Wang *et al*. (2022). Pour donner un aperçu concret du périmètre de la classification, voici les principales catégories d'exercices rencontrées : identification, classement, QCM, transformation, production, remise en ordre, oral, association, dictée et justification. Ces classes varient en fonction de l'unité linguistique mobilisée (mot, phrase, lettre, etc.) et du mode d'interaction (écriture, coche, échange, etc.).

– **Les adaptations** : une fois les exercices classifiés, ils sont transformés en formats interactifs HTML afin de faciliter l'interaction avec les élèves. Pour chaque classe d'exercice, un type d'adaptation spécifique est défini, en tenant compte à la fois de la nature de l'activité (QCM, dictée, association, etc.) et des capacités motrices, visuelles ou cognitives des élèves.

Par exemple, dans un exercice où l'élève doit « souligner le verbe », l'adaptation consistera à proposer une interaction par clic sur le mot correspondant, plutôt qu'une écriture manuscrite difficilement accessible pour certains élèves. De même, un exercice de classement pourra être adapté en glisser-déposer, tandis qu'un QCM utilisera des cases à cocher élargies.

Ces adaptations interactives prennent également la forme d'ajustements visuels (taille des caractères, espacement) ou de simplification du langage, et visent à maintenir les objectifs pédagogiques tout en respectant les capacités spécifiques des enfants.

Par ailleurs, une plateforme de vérification et un processus d'évaluation des exercices adaptés sont abordés dans (Pacini *et al.*, 2023).

3. Problématiques et hypothèses

Dans ce papier, nous nous concentrons sur l'extraction déjà réalisée, ses défis, et notre nouvelle approche d'extraction avec la vision par ordinateur.

Les méthodes d'extraction et de structuration précédemment développées par notre équipe, notamment dans les travaux de Lincker, Guinaudeau *et al.* (2023), reposent principalement sur des modèles de langue ou des architectures multimodales. Toutefois, ces approches présentent certaines limitations, en particulier une difficulté de généralisation entre différentes collections de manuels scolaires, en raison des variations de mise en page (voir Figure 4).

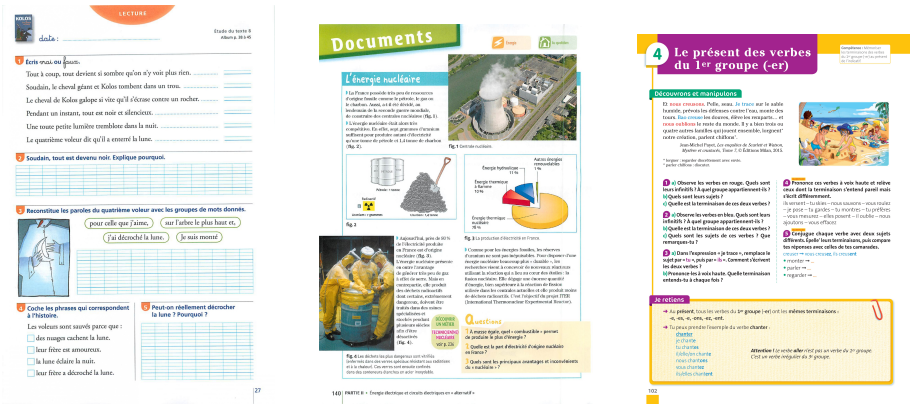
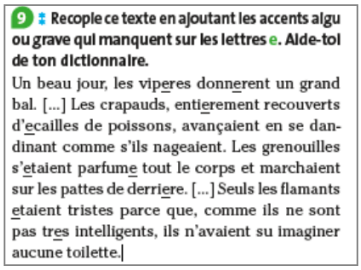


FIGURE 4 – Des pages provenant de différents manuels scolaires

Par ailleurs, l'annotation manuelle demeure une tâche longue et fastidieuse, et l'association précise entre les éléments visuels (tableaux, encadrés, pictogrammes, etc.) et leur contenu textuel reste complexe. De plus, la majorité des modèles existants sont principalement optimisés pour la langue anglaise, ce qui constitue un obstacle supplémentaire dans le contexte francophone.

Les exercices personnalisés avec des objectifs différents incluent des cas particuliers comme les fautes d'orthographe et de grammaire volontaires dans certains exercices (voir Figure 5). Ces fautes sont souvent introduites pour tester la capacité des élèves à identifier et corriger des erreurs, ajoutant une couche de complexité pour les modèles de langue et les systèmes de traitement de texte. En effet, ces derniers peuvent soit ne pas bien interpréter le texte, ce qui entraîne une classification et une structuration incorrectes, soit ne pas reconnaître ces erreurs comme intentionnelles et les corriger automatiquement à l'aide de leur dictionnaire intégré, altérant ainsi la nature pédagogique de l'exercice.



Texte sans accents

Absence d'espaces entre les mots

FIGURE 5 – Exemples d'exercices contenant des fautes volontaires

Extraire une consigne, un exemple ou un conseil est plus complexe que d'extraire simplement un titre ou un tableau, comme le faisaient la plupart des modèles entraînés sur des jeux de données tels que PubLayNet (Zhong *et al.*, 2019) ou DocBank (Li *et al.*, 2020), qui ciblent principalement des classes visuellement identifiables comme les titres, tableaux ou numéros de page. Les consignes et les exemples sont souvent intégrés dans des paragraphes plus larges et ne suivent pas toujours une structure typographique claire, rendant leur identification plus difficile. Un autre point bloquant est que ces modèles sont malheureusement efficaces uniquement avec les PDF natifs.

Pour surmonter ces limitations, il devient évident que l'analyse des images et des structures visuelles offre un moyen efficace de reconnaître divers éléments présents sur une page. Par exemple, lorsqu'un exercice contient deux listes verticales, il est facile et rapide d'identifier qu'il s'agit d'un exercice de type "liaison entre les choix".

Il apparaît également que, pour certaines catégories d'exercices, l'application de modèles de vision par ordinateur peut améliorer de manière significative la classification. En effet, de nombreux exercices présentent des éléments visuels caractéristiques — telles que des listes, des couleurs, des cases à cocher, des flèches ou des zones à remplir — qui permettent d'identifier leur type plus efficacement que par le seul traitement du texte. Ces indices visuels offrent une aide précieuse pour détecter la structure et la logique de l'exercice, en particulier lorsqu'ils suivent un format graphique récurrent.

Pour évaluer les performances de la vision par ordinateur, en particulier dans l'extraction et la structuration du contenu, nous avons testé plusieurs modèles tels que

LayoutParser (Shen *et al*., 2021), Detectron2 (Merz *et al*., 2023) et YOLO (Redmon *et al*., 2016). Ce dernier s'est avéré plus efficace pour notre cas d'utilisation, comme montré dans les résultats du (Tab.1), et a démontré une meilleure capacité à gérer les variations de mise en page et les éléments spécifiques aux manuels scolaires, notamment pour les documents scannés.

La suite de ce papier montre comment les méthodes utilisant la vision par ordinateur permettent de surmonter les limitations des approches existantes, notamment en ce qui concerne la gestion des documents scannés de qualité variable.

4. Méthodologie

Notre jeu de données est composé de 22 manuels scolaires français de l'école élémentaire, couvrant l'apprentissage de la langue, les mathématiques, les sciences et l'histoire-géographie. Ces manuels sont répartis équitablement en deux catégories : 11 manuels scannés et 11 manuels au format PDF natif. Cependant, notre concentration s'est principalement portée sur les manuels de langue.

Pour les PDF natifs, deux méthodes complémentaires d'annotation ont été mises en œuvre. Une première partie a été réalisée par des spécialistes de l'adaptation scolaire, notamment les membres de l'association Le Cartable Fantastique, via une plateforme développée en interne. Cette plateforme convertit les PDF en HTML, puis extrait automatiquement des éléments structurants à l'aide de règles basées sur les polices et la mise en page, après quelques annotations manuelles initiales des experts. Les annotations ainsi produites ont ensuite été converties au format *Labelme* [4] après transformation des pages en images.

Les pages restantes, non couvertes par cette plateforme, ainsi que les 11 manuels scannés, ont été directement convertés en images, puis annotés manuellement via l'outil *Labelme*. Chaque élément (consigne, titre, énoncé, etc.) a été encadré par une boîte rectangulaire. Pour les documents scannés, un prétraitement a été nécessaire (suppression des marges, correction des ombres, séparation des pages) afin d'améliorer la clarté avant l'annotation.

L'outil Labelme a ensuite été utilisé dans une boucle de rétroaction : après un certain taux d'annotation et un premier entraînement préliminaire (par exemple, dans l'annotation des exercices, nous avons entraîné le modèle de détection d'exercices), les prédictions du modèle ont été rouvertes dans Labelme pour correction. Cela nous a permis de construire progressivement un corpus annoté de grande taille, sans devoir tout annoter manuellement dès le départ.

Afin d'éviter le surapprentissage, 40 à 50 pages par manuel ont été sélectionnées pour l'entraînement, garantissant une diversité structurelle tout en limitant l'influence

4. https://labelme.io/

des formats spécifiques. La Figure (6) montre la répartition des annotations dans notre jeu de données.

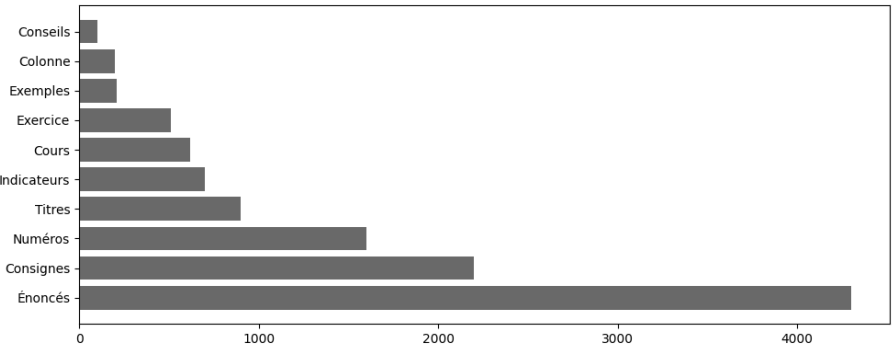


FIGURE 6 – Répartition des annotations dans notre jeu de données.

Le processus de modélisation des pages (extraction et structuration), illustré dans le diagramme d'activité comme deuxième étape (Figure 7), commence par la détection des exercices, où les zones pertinentes sont localisées sous forme de régions rectangulaires. Ensuite, la détection du format de page permet de distinguer les mises en page à une ou plusieurs colonnes. La détection des éléments de la page (tels que les consignes, énoncés, titres, indicateurs, exemples, etc.) est réalisée en fonction du format de la mise en page. Une fois ces éléments identifiés, une vérification est effectuée pour s'assurer qu'ils appartiennent bien à un exercice spécifique.

L'extraction de texte est ensuite réalisée à l'aide de *PDFAlto* [5] pour les fichiers PDF natifs, et de *RapidOCR* (RapidAI, 2022) pour les PDF scannés. Le texte extrait est alors associé aux éléments précédemment détectés, permettant une structuration cohérente et précise des contenus.

Enfin, les résultats sont organisés dans un fichier JSON structuré, selon un schéma prédéfini conforme aux spécifications de Attouche *et al*. (2024). Cette standardisation garantit la cohérence, l'interopérabilité et l'intégration aisée dans les systèmes éducatifs. Un exemple de sortie est illustré dans la figure 8, correspondant à l'analyse de l'exercice de la figure 1.

Une fois les exercices correctement extraits et structurés, la classification devient une étape plus simple et plus efficace. Dans le cadre du projet MALIN, 45 classes principales d'exercices ont été identifiées. Grâce à notre modèle, qui assure une extraction fiable des exercices, l'annotation manuelle, qu'elle soit dédiée à la classification ou à la vérification des classifications automatiques issues de Lincker, Guinaudeau *et al*. (2023), a été considérablement simplifiée : il suffit désormais de cliquer sur un exercice et de choisir une classe parmi les options proposées, comme illustré en Fig. 9.

5. https://github.com/kermitt2/pdfalto

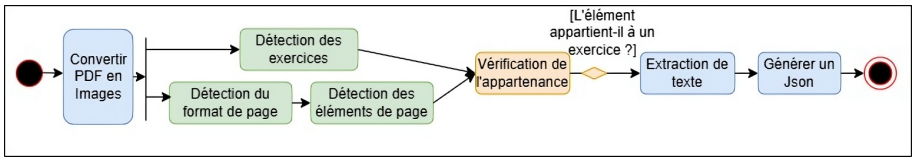


FIGURE 7 – Diagramme d'activité du processus d'extraction des exercices.


```
{ "manuel": "Outils pour le Français CE2 (2019)","isbn": "978-2-210-50538-4",
  "pages": [ { "id": "165",
      "exercices": [
        { "numéro": 5,
          "consigne": "Encadre le radical et souligne le suffixe des adjectifs suivants.",
          "énoncé": [ "a. peureux • pensif • excessif • malheureux",
           "b. maniable • craintif • lisible • lavable" ] }, ... ] }, ... ] }
```


FIGURE 8 – Sortie d'extraction et structuration.

Auparavant, cette tâche nécessitait soit une délimitation manuelle des exercices dans l'image, soit la copie intégrale du contenu textuel.

Ce processus garantit une extraction précise et adaptable du texte, couvrant tous les formats de manuels.

5. Résultats et Discussion

Nous débutons notre pipeline avec les modèles de reconnaissance du layout de la page et de localisation des exercices, en utilisant YOLOv10x sur des images de 736px avec un batch de 16 et l'optimiseur AdamW (Loshchilov, Hutter (2019)). Cette variante d'Adam (Adaptive Moment Estimation) inclut une correction de la régularisation par poids (weight decay), améliorant ainsi la stabilité et la convergence du modèle. Un taux d'apprentissage de 0,0001 avec une décroissance cosinus a également été appliqué. Les performances sont évaluées avec un seuil de confiance standard de 0,50 en vision par ordinateur. Cette configuration a permis d'atteindre une précision élevée pour les deux tâches principales : 97,8 % pour les exercices et 98 % pour le format de page. Cette étape assure une continuité fluide vers les suivantes.

Une fois le layout de la page et les exercices détectés avec précision, l'étape suivante consiste à identifier les différents éléments constitutifs des exercices.

La Table 1 compare les performances des trois modèles testés pour la détection des éléments : Detectron2, LayoutParser et YOLOv10x.

Detectron2 utilise un modèle GeneralizedRCNN avec un backbone ResNet-101 FPN (He *et al*. (2015) T.-Y. Lin *et al*. (2017)), un optimiseur SGD (Ruder (2017)) avec un taux d'apprentissage initial de 0.0025, et un entraînement sur des images de

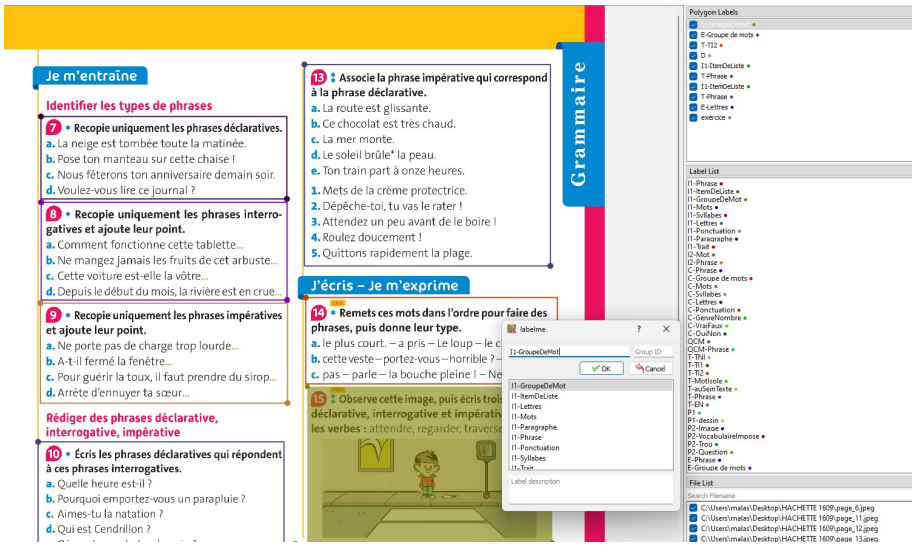


FIGURE 9 – Illustration de la méthode d'annotation pour la classification.

tailles variées (640 à 960 pixels) avec des augmentations de données incluant des flips horizontaux.

LayoutParser utilise également un modèle GeneralizedRCNN avec un backbone ResNet-101 FPN, un optimiseur SGD avec un taux d'apprentissage initial de 0.0025, et un entraînement sur des images de tailles variées (640 à 960 pixels) avec des augmentations de données incluant des flips horizontaux (Yang *et al*. (2023)).

YOLOv10x, avec les paramètres mentionnés précédemment.

Ce dernier, grâce à son approche en une seule passe, se distingue nettement en optimisant simultanément vitesse et précision. Avec un **AP50** de 78,1 % (précision moyenne à un seuil de recouvrement de 50 %) et un **AP (IoU 0.50:0.95)** de 55,6 % (moyenne des précisions sur plusieurs seuils, de 0,50 à 0,95), il surpasse largement les autres modèles. LayoutParser améliore légèrement les résultats de Detectron2, mais reste en retrait par rapport à YOLOv10x, notamment en raison de difficultés à générer correctement les boîtes englobantes des classes détectées. Detectron2 seul montre des performances plus limitées et inférieures aux autres modèles.

L'analyse par catégorie révèle que YOLOv10x excelle dans l'identification de la majorité des éléments structurants des documents, atteignant 92,2 % pour les numéros, 97,7 % pour les titres et 92,3 % pour les conseils. Cependant, ses performances sont plus faibles pour les "exemples" (53,8 %) en raison d'un nombre limité d'annotations (<100, comme montré dans Fig. 6), et pour les "cours" (41,8 %), car peu de pages contenant des cours ont été annotées, l'objectif principal étant l'extraction des exercices.

TABLEAU 1 – Performance Comparative des Modèles d'Analyse de Mise en Page

Métrique / Classe	detectron2	LayoutParser	YOLOv10x
AP50 (IoU 0.50)	0.396	0.540	**0.781**
AP (IoU 0.50:0.95)	0.187	0.262	**0.556**
consigne	0.227	0.235	**0.811**
énoncé	0.083	0.098	**0.837**
numéro	0.459	0.461	**0.922**
titre	0.248	0.298	**0.977**
indicateur	0.385	0.359	**0.824**
cours	0.072	0.063	**0.418**
exemple	0.017	0.108	**0.538**
conseil	0.001	0.475	**0.923**

Note : Les entraînements ont été réalisés pour la même durée et avec le même matériel.

Certaines erreurs observées sont principalement dues aux annotations, qui combinent des annotations générées automatiquement et manuelles, présentant des différences au niveau des marges et du padding des boîtes englobantes. Les annotations générées automatiquement sont généralement plus précises, tandis que les annotations manuelles introduisent parfois des marges additionnelles aléatoires.

6. Conclusion

Les contributions du projet résident dans la démonstration du potentiel de la vision par ordinateur pour comprendre des documents complexes comme les manuels scolaires, ainsi que dans l'obtention de résultats prometteurs pour l'extraction et la structuration de contenu multimodal, en particulier pour les documents numérisés.

Cette étude valide l'efficacité de la vision par ordinateur pour l'extraction et la structuration des contenus scolaires, avec une précision de 97,8 % pour les exercices et de 98 % pour le format de page. Les éléments clés, comme les numéros (92,2 %), les titres (97,7 %) et les conseils (92,3 %), sont bien identifiés. Nos futurs travaux viseront à améliorer les classes moins performantes afin d'assurer une accessibilité encore plus inclusive.

L'automatisation de l'extraction des exercices a également simplifié la classification en réduisant la charge d'annotation. L'assignation des classes aux exercices est désormais plus rapide, préparant efficacement l'étape suivante du pipeline d'adaptation.

Les perspectives futures incluent l'amélioration des modèles et l'augmentation des données par la génération d'exercices et de manuels, ce qui, outre son utilité pédagogique directe, permettrait de remédier aux déséquilibres des classes.

Par ailleurs, les manuels traités sont essentiellement des manuels de français. Les premières expériences sur d'autres matières semblent montrer que les manuels de

mathématiques du primaire, étant également assez réguliers, devraient pouvoir être traités de la même manière (la géométrie posant néanmoins des problèmes spécifiques, non pour l'extraction, mais en termes d'adaptation). La structure des manuels d'histoire-géographie, de sciences physiques ou de sciences de la vie, beaucoup moins régulière, avec de nombreux schémas, documents et illustrations croisés, semble encore plus complexe.

En termes d'extraction et de structuration de documents complexes, une autre piste intéressante est l'extension des méthodes à d'autres ressources dont la structure reste assez proche de celle des manuels, notamment les revues de vulgarisation scientifique. Nous envisageons par exemple, dans cette perspective, d'explorer le fonds documentaire du CNUM [6].

Nous visons également d'autres approches, telles que les Approches End-to-End et les Modèles Multimodaux de Grande Taille (VLMs) comme GPT-4 (OpenAI *et al.*, 2024), Qwen (Bai *et al.*, 2023), LLaMA (Touvron *et al.*, 2023) pour améliorer la précision et l'efficacité de l'extraction et de la structuration des contenus scolaires.

En conclusion, cette étude propose un pipeline d'analyse et de structuration de documents complexes. De plus, en contribuant significativement au processus d'adaptation automatisée des manuels scolaires, elle ouvre la voie à une éducation plus inclusive.

Remerciements

Ce travail a été soutenu par le projet ANR-21-CE38-0014 MALIN.

Bibliographie

Attouche L., Baazizi M.-A., Colazzo D., Ghelli G., Sartiani C., Scherzinger S. (2024, janvier). *Validation of modern json schema: Formalization and complexity* (vol. 8) n° POPL. New York, NY, USA, Association for Computing Machinery. Consulté sur https://doi.org/10.1145/3632891

Bai J., Bai S., Chu Y., Cui Z., Dang K., Deng X. *et al.* (2023). *Qwen technical report.* Consulté sur https://arxiv.org/abs/2309.16609

Devlin J., Chang M.-W., Lee K., Toutanova K. (2019). *Bert: Pre-training of deep bidirectional transformers for language understanding.* Consulté sur https://arxiv.org/abs/1810.04805

France. (2005). *Loi n° 2005-102 du 11 février 2005 pour l'égalité des droits et des chances, la participation et la citoyenneté des personnes handicapées.* Consulté sur https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000000809647 (Consulté le 2025-04-22)

France. (2013). *Loi n° 2013-595 du 8 juillet 2013 d'orientation et de programmation pour la refondation de l'école de la république.* Consulté sur https://www.legifrance.gouv.fr/dossierlegislatif/JORFDOLE000026973437/ (Consulté le 2025-04-22)

6. https://cnum.cnam.fr/

He K., Zhang X., Ren S., Sun J. (2015). *Deep residual learning for image recognition.* Consulté sur https://arxiv.org/abs/1512.03385

Le H., Vial L., Frej J., Segonne V., Coavoux M., Lecouteux B. *et al.* (2020). *Flaubert: Unsupervised language model pre-training for french.* Consulté sur https://arxiv.org/abs/1912.05372

Li M., Xu Y., Cui L., Huang S., Wei F., Li Z. *et al.* (2020). *Docbank: A benchmark dataset for document layout analysis.* Consulté sur https://arxiv.org/abs/2006.01038

Lin J., Yin H., Ping W., Lu Y., Molchanov P., Tao A. *et al.* (2024). *Vila: On pre-training for visual language models.* Consulté sur https://arxiv.org/abs/2312.07533

Lin T.-Y., Dollár P., Girshick R., He K., Hariharan B., Belongie S. (2017). *Feature pyramid networks for object detection.* Consulté sur https://arxiv.org/abs/1612.03144

Lincker E., Guinaudeau C., Pons O., Barbet I., Dupire J., Hudelot C. *et al.* (2023, juin). Classification automatique de données déséquilibrées et bruitées : application aux exercices de manuels scolaires. In C. Servan, A. Vilnat (Eds.), *Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 4 : articles déjà soumis ou acceptés en conférence internationale*, vol. 4, p. 121-130. Paris, France, ATALA. Consulté sur https://hal.science/hal-04130220

Lincker E., Pons O., Guinaudeau C., Barbet I., Dupire J., Hudelot C. *et al.* (2023). Layout- and activity-based textbook modeling for automatic pdf textbook extraction. In *Proceedings of the intelligent textbooks workshop at the 24th international conference on artificial intelligence in education (aied)*, p. 37–53. CEUR Workshop Proceedings. Consulté sur https://hal.science/hal-04184895 (Available under a Creative Commons Attribution 4.0 International License)

Loshchilov I., Hutter F. (2019). *Decoupled weight decay regularization.* Consulté sur https://arxiv.org/abs/1711.05101

Magnard. (2019). *Outils pour le français ce2 (2019) - manuel élève*. Magnard. Consulté sur https://www.magnard.fr/livre/9782210505384-outils-pour-le-francais-ce2-2019-manuel-eleve (Consulté en avril 2025)

Martin L., Muller B., Ortiz Suárez P. J., Dupont Y., Romary L., Clergerie de la *et al.* (2020). Camembert: a tasty french language model. In *Proceedings of the 58th annual meeting of the association for computational linguistics.* Association for Computational Linguistics. Consulté sur http://dx.doi.org/10.18653/v1/2020.acl-main.645

Merz G., Liu Y., Burke C. J., Aleo P. D., Liu X., Carrasco Kind M. *et al.* (2023, septembre). *Detection, instance segmentation, and classification for astronomical surveys with deep learning (deepdisc): Detectron2 implementation and demonstration with hyper suprime-cam data* (vol. 526) n° 1. Oxford University Press (OUP). Consulté sur http://dx.doi.org/10.1093/mnras/stad2785

OpenAI, Achiam J., Adler S., Agarwal S., Ahmad L., Akkaya I. *et al.* (2024). Gpt-4 technical report. Consulté sur https://arxiv.org/abs/2303.08774

Pacini L., Dupire J., Barbet I., Pons O., Guinaudeau C., Mousseau V. *et al.* (2023). Textbook accessibility for children with dyspraxia and visual disabilities. In *17th international conference of the association for the advancement of assistive technology in europe (aaate 2023).*

Rahtz S., Walsh N., Burnard L. (2004). A unified model for text markup: Tei, docbook, and beyond. In *Proceedings of xml europe.*

RapidAI. (2022). *Rapidocr.* https://github.com/RapidAI/RapidOCR. (Disponible sur https://github.com/RapidAI/RapidOCR)

Redmon J., Divvala S., Girshick R., Farhadi A. (2016). *You only look once: Unified, real-time object detection.*

Ruder S. (2017). *An overview of gradient descent optimization algorithms.* Consulté sur https://arxiv.org/abs/1609.04747

Shen Z., Zhang R., Dell M., Lee B. C. G., Carlson J., Li W. (2021). *Layoutparser: A unified toolkit for deep learning based document image analysis.* Consulté sur https://arxiv.org/abs/2103.15348

Stahn L.-L., Hennicke S., De Luca E. W. (2016). Using tei for textbook research. In *Proceedings of the workshop on language technology resources and tools for digital humanities (lt4dh).*

Touvron H., Lavril T., Izacard G., Martinet X., Lachaux M.-A., Lacroix T. *et al.* (2023). *Llama: Open and efficient foundation language models.* (vol. abs/2302.13971). Consulté sur http://dblp.uni-trier.de/db/journals/corr/corr2302.html#abs-2302-13971

Walsh N., Hamilton R. L. (2010). *Docbook 5: The definitive guide: The official documentation for docbook.* " O'Reilly Media, Inc.".

Wang J., Jin L., Ding K. (2022). *Lilt: A simple yet effective language-independent layout transformer for structured document understanding.* Consulté sur https://arxiv.org/abs/2202.13669

Xu Y., Li M., Cui L., Huang S., Wei F., Zhou M. (2020, août). Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th acm sigkdd international conference on knowledge discovery amp; data mining*, p. 1192–1200. ACM. Consulté sur http://dx.doi.org/10.1145/3394486.3403172

Xu Y., Xu Y., Lv T., Cui L., Wei F., Wang G. *et al.* (2022). *Layoutlmv2: Multi-modal pre-training for visually-rich document understanding.* Consulté sur https://arxiv.org/abs/2012.14740

Yang S., Xiao W., Zhang M., Guo S., Zhao J., Shen F. (2023). *Image data augmentation for deep learning: A survey.* Consulté sur https://arxiv.org/abs/2204.08610

Zhong X., Tang J., Yepes A. J. (2019). *Publaynet: largest dataset ever for document layout analysis.* Consulté sur https://arxiv.org/abs/1908.07836

L'architecture d'entreprise au service de la transformation numérique

Sarah Triki [1], **Christophe Ponsard** [1], **Mounir Touzani** [2]

1. *University of Namur, Belgique*
   sarah.triki07@gmail.com,christophe.ponsard@unamur.be
2. *Chercheur indépendant, France*
   mounir.touzani@inrae.fr

RÉSUMÉ. *La transformation numérique (TN) permet d'introduire des innovations technologiques au sein des entreprises et de faire émerger de nouveaux modèles d'affaires. Ceci nécessite cependant une évolution de l'organisation et de ses processus métier. L'architecture d'entreprise (AE) malgré qu'elle soit originellement plus orientée IT s'aligne de plus en plus sur la stratégie métier via des approches collaboratives et agiles. Dans ce contexte, cet article explore comment l'AE apporte un soutien à la démarche de la TN d'une organisation. Ce travail se base sur un état de l'art de la littérature afin d'identifier les cadres structurant les plus utilisés et méthodes d'évaluation de maturité. Sur cette base, nous avons identifié des facteurs de renforcement ainsi que des barrières, et avons formulé plusieurs hypothèses. Celles-ci ont été validées à l'aide d'une enquête en entreprise dont les résultats ont été croisés avec des travaux similaires. Nos constats ont mené à des recommandations pour un alignement optimal de l'AE avec son processus de TN.*

ABSTRACT. *Digital transformation makes it possible to introduce technological innovations within companies while also giving rise to new business models. However, this requires an evolution of the organization and its business processes. Although originally more IT-oriented, enterprise architecture is increasingly aligning with business strategy through collaborative and agile approaches. In this context, this article explores how enterprise architecture supports an organization's digital transformation process. The work is based on a state-of-the-art review of the literature in order to identify the most commonly used structuring frameworks and maturity assessment methods. Based on this, we identified reinforcing factors and barriers and formulated hypotheses. These hypotheses were validated through a corporate survey whose results were compared with similar studies. Our findings led to recommendations for an optimal alignment of the enterprise architecture with its digital transformation process.*

MOTS-CLÉS : *Entreprise architecture, digital transformation, digital maturity, survey, recommendations*

KEYWORDS: *Architecture d'entreprise, transformation numérique, maturité digitale, enquête, recommandations*

1. Introduction

Pour rester concurrentielles dans un marché mondialisé en constante évolution, les entreprises doivent s'adapter et innover en permanence en tirant parti des technologies émergentes. La Transformation Numérique (TN) des entreprises résulte de changements impliquant « l'utilisation d'artefacts, de systèmes et de symboles numériques au sein de l'entreprise et dans son environnement » (Bounfour, 2016). Cette transformation impacte divers aspects de l'entreprise, notamment son organisation, ses processus métier, ses systèmes d'information et son infrastructure, formant ensemble une Architecture d'Entreprise (AE). Le potentiel du numérique pousse les entreprises à repenser leur création de valeur. Cependant, beaucoup d'entreprises restent sans stratégie numérique claire pour leur transformation. Bien que des cadres d'AE existent pour guider le changement, leur usage reste limité (Goerzig, Bauernhansl, 2018).

La conduite de la TN est un processus complexe qui requiert l'utilisation de méthodes et d'outils appropriés. L'AE offre un cadre pour orienter l'évolution de l'organisation et de son système d'information. Cela implique l'élaboration d'une stratégie à long terme, englobant des objectifs pluriannuels, une planification des activités et une anticipation des besoins en personnel. En fournissant ce cadre holistique avec des pratiques structurées garantes de la cohérence des évolutions et facilitant la gouvernance, l'AE peut constituer un levier stratégique et opérationnel permettant de réussir la transformation numérique (ITU, 2019).

Dans ce contexte, la question de recherche centrale de notre article se formule ainsi : « Comment une démarche d'AE peut-elle soutenir efficacement la TN d'une organisation ? »

Pour y répondre, cet article suit une structure méthodique. La section 2 propose une revue de la littérature, offrant une compréhension approfondie de la TN et de l'AE. Elle se conclut par une synthèse offrant une première réponse identifiant les barrières et facteurs d'adoption des TN et permettant de mesurer l'apport de l'AE. Notre question de recherche est alors affinée en plusieurs hypothèses auxquelles la section 3 répond sur la base d'une enquête réalisée auprès d'entreprises. Afin de confirmer nos constats, ceux-ci sont comparés avec des résultats d'autres enquêtes dans la section 4. Enfin, la section 5 conclut en dégageant des pistes pour nos travaux futurs.

2. État de l'art

Après avoir posé les définitions de la TN et de l'AE, cette section identifie les modèles, leviers et barrières pour les articuler.

2.1. Définitions

Transformation Numérique (TN). De nombreuses définitions ont été proposées dans la littérature (Schallmo, Williams, 2017)(Ziyadin *et al.*, 2020). Le but est surtout d'utiliser de nouvelles technologies très innovantes pour travailler plus efficacement

et accroître la productivité, la création de valeur et le bien-être social (Ebert, Duarte, 2018). Ceci engendre un changement profond du fonctionnement de l'organisation se traduisant par un modèle opérationnel et commercial fondé sur l'exploitation des données et des réseaux (Mergel, al., 2019).

Architecture d'Entreprise (AE). L'AE peut se définir comme une démarche d'alignement stratégique du système d'information, ancrée dans les enjeux et objectifs de l'entreprise (CIGREF, 2008). Cette démarche illustre bien la dimension de transformation et s'appuie sur une modélisation du fonctionnement de l'organisation décrivant la structure et le comportement de ses processus, de ses systèmes d'information, de son personnel et de ses unités organisationnelles, de manière de réaliser cet alignement (Office québécois de la langue française, 2007).

2.2. Transformation numérique

Dans ce volet de l'état de l'art dédié à la transformation numérique, nous identifierons les principales barrières et facteurs d'adoption afin d'examiner ultérieurement si des actions peuvent être bénéfiques au niveau de l'architecture d'entreprise. Nous n'entrerons pas ici dans une revue générale disponible dans (Zaoui, Souissi, 2020) ni dans les différents modèles d'affaires numériques disponibles (Weill, Woerner, 2013)(Winer, Bock, 2017), afin de nous concentrer sur l'évaluation de la maturité qui nous sera utile pour mener notre enquête.

2.2.1. Barrières à lever

Les entreprises font face à des défis majeurs dans leur transformation numérique. Ces défis concernent les entreprises de toutes tailles et plus spécifiquement les PME. Ainsi, en Belgique, 66% des PME interrogées ont signalé rencontrer des obstacles entravant leur processus de digitalisation (SPF Economie, 2024). Plusieurs taxonomies de barrières sont données dans (Jones *et al.*, 2021). Nous en identifions ici les principales qui sont pertinentes pour notre question de recherche.

– **Les ressources financières et humaines limitées** peuvent freiner les investissements nécessaires à l'identification, à l'acquisition et à l'intégration de solutions numériques nécessaires à la TN, y compris les dépenses connexes de formation, et d'innovation. Ceci concerne plus spécifiquement les PME et leur impose d'être prudentes dans la sélection et dans la mise en œuvre des technologies numériques (OECD, 2017)(SPF Economie, 2024), à l'exception notable des startups technologiques. Certains domaines comme l'Industrie 4.0 exigent aussi des investissements conséquents retardant la TN des PME de ce domaine (Faller, Feldmüller, 2015).

– **Le manque de compétences** est un obstacle majeur à la digitalisation des entreprises dans de nombreux domaines. En particulier, ceci concerne plus de 50% des entreprises manufacturières (Jones *et al.*, 2021). Les PME sont particulièrement concernées, notamment 78% des PME Wallonnes (Digital Wallonia, 2022), et sont en outre souvent accaparées par les tâches quotidiennes, limitant leur disponibilité pour le développement de nouvelles solutions (Goerzig, Bauernhansl, 2018).

– **La résistance au changement** entrave la digitalisation des entreprises de toute taille. Elle est l'un des 3 obstacles clefs signalé par (Sailer *et al.*, 2019) ainsi que dans le domaine manufacturier (Jones *et al.*, 2021). Elle concerne aussi les PME et a été rapportée dans les contextes français (Peillon, Dubruc, 2019) et belge, avec environ 25% des entreprises reconnaissant ce problème (SPF Economie, 2024). Cependant ce facteur est souvent sous-estimé et non reconnu par les entreprises (Leipzig *et al.*, 2017).

– **L'absence de stratégie de transformation numérique** est un facteur pointé par de nombreuses études résumées par (Mahmood *et al.*, 2019). En effet, la simple adoption de technologies ne suffit pas : la clé réside dans la manière dont les technologies numériques sont intégrées et exploitées pour transformer l'activité de l'entreprise (Peillon, Dubruc, 2019). Concernant les PME, la majorité ne disposent pas d'une stratégie claire et les actions numériques sont souvent ponctuelles, intuitives, et peu coordonnées (Bouncken, Schmitt, 2022).

– **L'adoption des technologies numériques** constitue en soi un socle de la TN mais peut constituer un mur technologique difficile pour des secteurs moins numérisés tels que la construction ou l'agriculture (Manyika *et al.*, 2015). D'autres obstacles comprennent l'interdépendance des technologies, la nécessité de repenser la conception et les enjeux de cybersécurité (Peillon, Dubruc, 2019)(Vogelsang *et al.*, 2019).

2.2.2. Leviers d'adoption

Divers facteurs mis en évidence par la littérature, favorisent l'intégration des technologies numériques au sein des entreprises, notamment la culture organisationnelle, le leadership, l'expérience client, la technologie et l'agilité.

– **La culture organisationnelle** joue un rôle clé dans la TN en améliorant la créativité, le contrôle et la performance. Il est crucial de promouvoir une culture favorisant l'innovation, l'amélioration continue et l'orientation client (Gamache *et al.*, 2020). Elle transforme les opérations, renforce la collaboration et l'apprentissage continu (Jonathan *et al.*, 2021)(Cantemir *et al.*, 2023). Elle permet aussi d'améliorer les interactions avec les clients et les employés, tout en développant l'adaptabilité et la communication interne. En outre, une culture qui encourage la participation active des employés et managers facilite le changement et sa gestion efficace (Ozguner, 2021).

– **Le leadership** est essentiel pour orienter une entreprise face aux évolutions du marché (Gamache *et al.*, 2020). Par leur soutien, les dirigeants ancrent les valeurs numériques dans la culture d'entreprise, favorisent l'innovation et l'esprit entrepreneurial (Holotiuk, Beimborn, 2017). Une TN réussie repose sur un leadership fort, une vision claire et des ressources adaptées (Cantemir *et al.*, 2023).

– **Le rôle Chief Digital Officers (CDO)** permet d'incarner la TN et de mobiliser ses compétences pour aligner la stratégie sur les valeurs de l'entreprise. Le CDO doit convaincre les parties prenantes de l'importance du digital et coordonner les actions internes. La DSI (Direction des Systèmes d'Information), souvent concentrée sur l'opération et la maintenance des outils, ne peut assumer seule cette mission (Ducrey, Vivier, 2017)(Dudézert, 2018).

– **L'expérience client** est au cœur de la TN. L'orientation client et l'amélioration des interactions client sont des motivations clés pour la digitalisation (Mhlungu *et al.*, 2019). La TN elle-même peut être déclenchée par l'évolution des attentes des clients (Osmundsen *et al.*, 2018). Les entreprises s'améliorent particulièrement dans le suivi des relations client et la création de services plus conviviaux (Aghakhani *et al.*, 2021).

– **La technologie** est un catalyseur clé de la TN, reposant sur des outils à présent communs tels que les CRM, ERP et infrastructures réseau. L'IoT et le Big Data sont déjà largement adoptés (Cantemir *et al.*, 2023) tandis que d'autres technologies comme l'IA offrent de nombreuses opportunités (Avasarala, 2020).

– **L'agilité** est essentielle pour réussir la TN. Selon McKinsey, les entreprises adoptant des pratiques agiles ont deux fois plus de chance de surpasser leurs attentes de performance dans leur TN (Bughin *et al.*, 2019). Un leadership agile favorise l'adaptabilité et la flexibilité organisationnelle, accélérant le changement (Ozguner, 2021) tandis que l'agilité informatique facilite sa mise en œuvre (Fuchs, Hess, 2018).

2.2.3. Modèle de maturité de la TN

Il existe de nombreux indicateurs de maturité surtout développés par des cabinets de consultances pour les grandes entreprises. Nous en donnons ici un aperçu :

– Le modèle de Deloitte (Deloitte, 2018) analyse la maturité numérique via cinq dimensions : client (expérience et engagement), stratégie (avantage concurrentiel par le numérique), technologie (gestion et sécurisation des données), opérations (gestion agile et automatisation) et organisation/culture (gouvernance et gestion des talents). Il propose cinq niveaux de maturité, depuis « Initiation » jusqu'à « Leader ».

– Le modèle Forrester (Bounfour, 2016) repose sur quatre dimensions : culture (compétences numériques et formation), technologie (adoption des innovations), organisation (stratégie et gouvernance), et perspectives (analyse des performances à partir des données). Il classe les entreprises de manière croissante en sceptiques, adoptants, collaborateurs et différenciateurs.

– Le modèle MIT/Capgemini (Westerman *et al.*, 2012) identifie quatre catégories selon leur maturité numérique : débutants (adoption limitée du numérique), fashionistas (technologies utilisées sans cohérence stratégique), conservateurs (stratégie rigide freinant l'innovation) et digiratis (maîtrise avancée et vision stratégique forte).

– Le DigiScore (Digital Wallonia, 2020) évalue la maturité numérique des entreprises wallonnes sur base d'une centaine de critères regroupés selon quatre axes : infrastructure (utilisation des outils technologiques), organisation (gestion de projets et culture du numérique), processus (intégration du numérique pour optimiser les flux d'information), et stratégie (intégration des technologies dans la stratégie). Les DigiScores sont analysés dans le baromètre digital environ tous les 2 ans.

– La matrice du Hub Institute (Ducrey, Vivier, 2017) repose sur six chantiers : « Leadership », « Culture & Organisation », « Technologies », « Données », « Exp. Clients & Marketing 2.0 » (expérience client) et « Mesure » (mesure de la performance). Pour chaque chantier, on passe par plusieurs étapes : d'abord préparer le chan-

gement (« Auditer », « Planifier » et « Tester »), puis le mettre en œuvre (« Déployer » et « Optimiser »). L'ensemble de ces étapes est résumé dans la matrice présentée à la figure 1. Cette matrice est couplée avec cinq niveaux de maturité, du niveau 0 (absence de pratiques numériques) au niveau 4 (utilisation des données en temps réel et automatisation avancée) qui sont moyennés sur les 6 chantiers (Gamache *et al.*, 2020).

	Auditer	Planifier	Tester	Déployer	Optimiser
Leadership	Compréhension & Vision	Stratégie & New Business Model	Pilote & Lab	Roadmap & Soutien/ exemplarité	Itération & optimisation
Culture et organisation	QI Digital & Acculturation	Compétences & Organisation	Process & Policies	Formations & Changement	Collaboration
Technologies	Ecosystème & Architectures	Intégration Business & IT	Everywhere Commerce	IOT, Cloud, SAAS, API	Agilité & Open Innovation
Données	Mapping datas existantes	Unification CRM & DMP	Légal & Cybersécurité	Data-visualisation & Datamining	Big Data & Prédictif
Exp. Clients & Marketing 2.0	Connaissance Clients	Contenus, Services & Exp	Social, Mobile, Vidéo	Engagement & Mediaplanning Omnicanal	Temps Réel
Mesure	Bonnes pratiques	KPIS	Analytics & Metrics	Tableaux de bord	Benchmarks

FIGURE 1. *matrice des 6 chantiers du HUB Institute*

2.3. Architecture d'entreprise

Nous donnons ici un aperçu des principaux frameworks répandus pour l'EA et la manière dont ils soutiennent la TN :

– **TOGAF (The Open Group Architecture Framework)** (The Open Group, 2018) est un cadre d'architecture structurant la création, l'évolution et la gestion des architectures d'entreprise. Il repose sur l'Architecture Development Method (ADM), un processus itératif structuré en phases permettant de définir une architecture cible alignée avec les objectifs métier et IT. TOGAF fournit des principes, des modèles et de bonnes pratiques pour concevoir une architecture optimisée et évolutive. **Concernant la TN**, TOGAF facilite la digitalisation en structurant l'adoption de nouvelles technologies (cloud, big data, IA) et en assurant une migration progressive vers des systèmes modernes et flexibles. Il permet de répondre aux défis d'innovation et d'alignement stratégique en gérant les transformations de manière cohérente et agile.

– **Le Framework de Zachman** (Zachman, 2011) repose sur une matrice 6x6 combinant perspectives organisationnelles et dimensions analytiques (Quoi, Comment, Qui, Où, Quand, Pourquoi). Il permet une classification rigoureuse des composants d'une organisation et facilite leur modélisation systématique, garantissant une vision complète et cohérente de l'AE. **Concernant la TN**, Zachman la structure en alignant les acteurs de l'entreprise sur une vision commune. Il facilite la transition vers des modèles digitaux en structurant la modernisation des processus et en intégrant des outils comme UML pour la conception des systèmes d'information.

– **COBIT (Control Objectives for Information and Related Technology)** (ISACA, 2012) est un cadre de gouvernance IT développé par l'ISACA qui définit

des processus et pratiques permettant de gérer et d'optimiser l'IT au service des objectifs stratégiques. Il est structuré en deux domaines : la gouvernance et la gestion, et couvre des processus liés à la gestion des ressources, des risques et de la performance IT. **Concernant la TN**, COBIT garantit que les initiatives de transformation digitale sont gouvernées efficacement en équilibrant valeur, risque et performance IT. Il permet d'aligner la digitalisation avec les exigences réglementaires et d'optimiser l'utilisation des ressources IT pour assurer une transformation sécurisée et conforme.

– **IT4IT** (The Open Group, 2024) est un cadre d'architecture développé par l'Open Group qui vise à gérer l'ensemble du cycle de vie des services et produits numériques. Il repose sur une architecture de référence divisée en flux de valeurs (de la stratégie à l'exploitation), assurant une gestion optimisée des solutions IT et une meilleure intégration des processus digitaux. **Concernant la TN**, IT4IT soutient l'adoption des approches DevOps et automatisation des services, rendant les organisations plus agiles et réactives face aux évolutions technologiques. Il facilite l'intégration des outils cloud et l'orchestration des processus numériques.

2.4. Synthèse sur l'apport de l'AE à la TN depuis l'état de l'art

Intrinsèquement, l'AE a pour but d'accompagner une transformation, donc elle a un rôle clé à jouer dans la transformation numérique, à différentes étapes.

– en amont, l'AE permet de capturer le business model digital et une vision claire de l'entreprise afin d'intégrer de nouveaux services numériques dans la stratégie globale. L'AE permet ici d'évaluer la situation et d'identifier les lacunes.

– ensuite, pour la planification de la TN, l'AE capture sa feuille de route et permet d'aligner les technologies avec les besoins en tenant compte des risques. Ceci est un facteur de succès des projets numériques souligné dans (Jonathan *et al.*, 2023).

– en préparation et en soutien à la réalisation, l'AE intègre la gestion du changement y compris des dimensions de formation et de communication. L'AE favorise aussi l'agilité organisationnelle et l'intégration des partenaires et des écosystèmes.

– au niveau de la réalisation, l'AE offre un suivi précis avec des indicateurs clés de performance qui facilitent l'évaluation continue.

– enfin, en matière de sécurité, elle prend en compte la protection des données et la gestion des identités.

Afin d'estimer l'apport de l'AE de manière plus précise et « couvrir » le processus de la TN, nous avons utilisé le modèle matriciel du Hub Institute présenté en figure 1 pour produire la figure 2 reprenant des ses cellules les leviers résumés ci-dessus (en vert) et d'autres plus larges (en bleu). On voit que l'AE joue un rôle dans de nombreux facteurs technologiques mais aussi au niveau de l'accompagnement du leadership. Sans surprise, elle soutient intégralement la colonne de planification. Chaque activité peut ensuite être analysée sous l'angle de leviers ou d'obstacles à lever, par exemple : la définition d'une stratégie TN sur la vision et le business model AE, la mise en place de compétences pour les capacités, ou encore la gestion du changement.

	Auditer	Planifier	Tester	Déployer	Optimiser
Leadership	Vision	Business model	Projet pilote	Roadmap	
Culture et org.	Culture	Capacités	Processus	Formation Changement	Collaboration
Technologies	Ecosystème Architecture	Intégration Business/IT		Solutions	Agilité
Données	Inventaires	Unification	Réglementation Cybersécurité		
Exp. Client	Attentes client	Services		Communication	
Mesure	Bonnes pratiques	Collecte KPI			

Légende

Leviers TN
Autres leviers

FIGURE 2. *Positionnement de l'AE dans la feuille de route du Hub Institute.*

3. Enquête sur l'état des pratiques

A ce stade, il est utile d'affiner notre question de recherche « Comment une démarche d'AE peut-elle soutenir efficacement la TN d'une organisation » en formulant des hypothèses plus précises auxquelles nous allons tenter de répondre à l'aide d'une collecte de données auprès d'entreprises.

3.1. Formulation d'hypothèses à valider auprès des entreprises

Nous avons choisi de traiter trois hypothèses liées à l'alignement entre la stratégie et la technologie, aux caractéristiques de mise en œuvre de l'AE, ainsi qu'à l'influence mutuelle entre la maturité en matière d'AE et de TN :

– Hypothèse 1 : Les entreprises veillent à ce que les technologies digitales soient alignées avec leur métier dans le cadre de leur TN en s'appuyant sur l'AE.

– Hypothèse 2 : Les entreprises qui surmontent les défis liés à la transformation numérique ont une AE mieux adaptée et plus agile.

– Hypothèse 3 : L'efficacité de la TN dans une entreprise est directement liée à la qualité de son AE.

3.2. Conception et distribution du questionnaire

Afin de réaliser un questionnaire d'enquête, nous nous sommes appuyés sur des études de pratiques existantes en matière d'AE par rapport à la TN qui ont été identifiées dans notre état de l'art (Deloitte, 2018)(Gamache *et al.*, 2020)(Digital Wallonia, 2022). Ceci facilite aussi une analyse comparative des résultats réalisés à la fin de l'article. L'enquête a été structurée en 4 sections :

– **La caractérisation de l'entreprise** identifie le secteur d'activité (14 domaines), la taille (TPE, PME, grande entreprise) et son ancienneté (moins de 5 ans, 5-10 ans, plus de 10 ans). Elle détermine aussi le profil du répondant (CEO, CIO, CDO).

– **La TN** identifie les activités en cours et les domaines concernés, examine les pratiques et obstacles rencontrés, analyse les leviers mobilisés et mesure le niveau de maturité.

– **L'AE** est évaluée au niveau de la familiarité avec les frameworks d'AE, les objectifs (alignement IT-métier), les activités réalisées (modèles d'affaires, conception IT) et les approches/outils déployés en soutien.

– **La combinaison AE et TN** est évaluée en demandant d'évaluer l'utilité d'outils AE pour la TN, d'identifier des éléments clés de l'AE ainsi que des obstacles à la mise en œuvre d'outils de type AE pour la TN.

La formulation des questions se base essentiellement sur des choix multiples, avec plusieurs réponses possibles et en lien avec les choix identifiés dans l'état de l'art. Pour ne pas exclure d'autres réponses, une option ouverte « Autre » est souvent proposée. Le formulaire est disponible en ligne (Triki, Ponsard, 2024).

3.3. Caractéristiques générales de l'échantillon collecté

Nous avons collecté 18 réponses valides avec un panel représentatif des secteurs d'activités des entreprises, dont l'horeca, le numérique, la culture, la finance, les services aux entreprises, la logistique, la construction, l'agriculture et les soins de santé. En ce qui concerne la taille des entreprises, les PME ont représenté 56% des répondants, tandis que les grandes entreprises ont contribué à hauteur de 44%. En ce qui concerne la durée d'existence des entreprises : 67% pour les entreprises de plus de 10 ans, 5% pour celles entre 5 et 10 ans, et 28% pour celles de moins de 5 ans. La majorité (69%) était engagée dans un processus de TN, 25% en phase de réflexion stratégique et une minorité (6%) inactives en la matière.

3.4. Réponses aux hypothèses formulées

3.4.1. Hypothèse 1 - Les entreprises veillent à ce que les technologies digitales soient alignées avec leur métier dans le cadre de leur TN en s'appuyant sur l'AE

La figure 3 montre que l'objectif principal de l'AE est d'aligner les technologies de l'information sur les besoins métiers de l'entreprise, avec un taux pondéré de plus de 50%, largement supérieur aux autres objectifs. Cet objectif est systématiquement mentionné par les grandes entreprises (GE) et par 80% des PME.

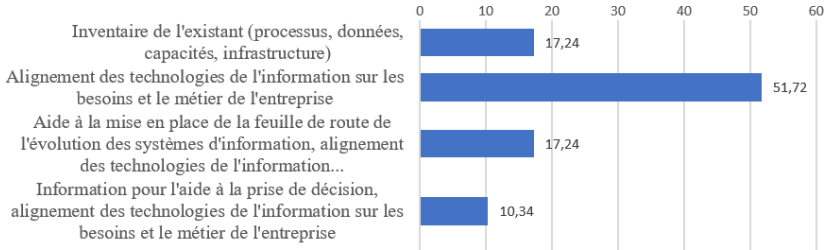


FIGURE 3. *Objectifs pour la mise en œuvre d'une AE*

Une analyse plus indirecte peut aussi se faire à la lumière des outils d'analyse pour un projet TN. Les deux outils les plus utilisés sont le diagramme de l'infrastruc-

ture IT (59% global, 50% PME, 70% GE) , suivi par l'analyse SWOT (41% global, 30% PME, 57% GE) et le modèle de processus métiers (30% global, 20% PME, 43% GE). Ces outils visuels facilitent la compréhension et la communication de la stratégie numérique et montrent que des bases sont présentes pour réaliser l'alignement entre les objectifs commerciaux et la mise en œuvre technologique au sein de l'AE. Sans surprise l'adoption est plus importante dans les GE. Les PME sont aussi plus ignorantes: environ 40% des outils en moyenne contre 10% pour les GE qui sont plus en phase d'analyse d'intérêt de certains outils. Les éléments collectés confirment donc bien cette première hypothèse.

3.4.2. Hypothèse 2 - Les entreprises qui surmontent les défis liés à la transformation numérique ont une architecture d'entreprise mieux adaptée et plus agile

Les principaux obstacles mentionnés par les entreprises ont été passés en revue. La figure 4 illustre que ceux-ci concernent surtout la complexité technique, le manque de compétences/ressources et la complexité des frameworks. Pour chaque entreprise nous avons examiné le spectre des problèmes rapportés et croisé l'information avec la quantité de pratiques AE déclarées notamment via leurs outils.

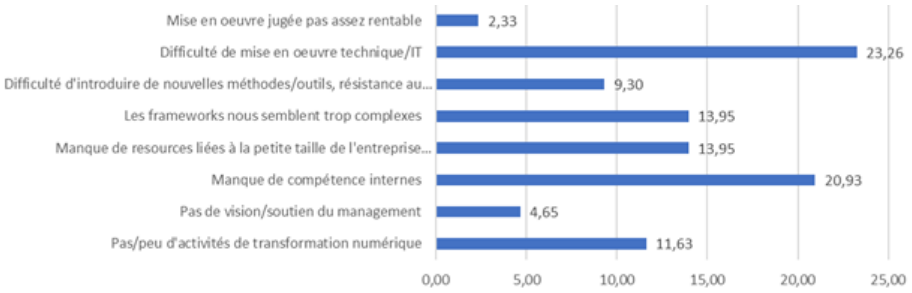


FIGURE 4. *Obstacles à la mise en œuvre d'outils aidant à la TN*

Concernant l'agilité, des indicateurs sont collectés via deux questions. Environ 60% des entreprises (GE et PME) le considèrent comme levier important. Dans les faits, ces pratiques sont effectivement en place dans 55% des entreprises (50% PME et 63% GE). Ce qui indique que le levier semble effectif. Les données collectées ne montrent pas cependant pas de lien significatif entre les entreprises qui ont adopté ces pratiques et le nombre de problèmes rapportés.

3.4.3. Hypothèse 3 - L'efficacité de la transformation numérique dans une entreprise est directement liée à la qualité de son architecture d'entreprise

Pour évaluer l'hypothèse, nous avons réalisé une analyse corrélant les maturités en TN et AE. La méthodologie d'évaluation repose sur une évaluation subjective menée dans le questionnaire sur la base des indicateurs (simplifiés) suivants :

– La maturité TN est évaluée via le niveau d'avancement déclaré : début de réflexion (1), conception de projet (2), mise en œuvre (3), évaluation (4).

– La maturité AE est évaluée via le type de soutien : pas explicite/organisé (1), par projet ponctuel (2), via une approche sur plusieurs projets (3), via un cadre général (4).

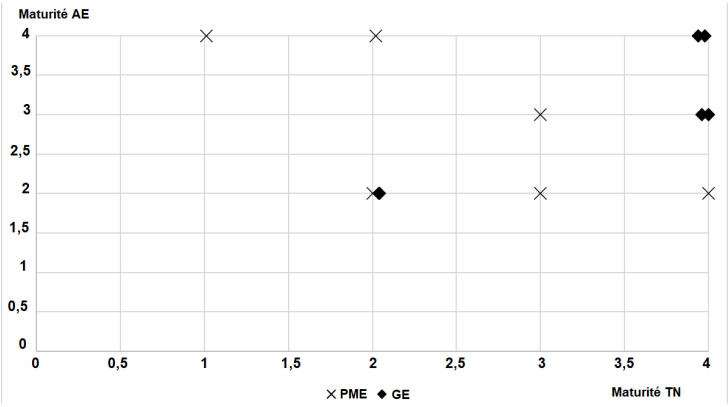


FIGURE 5. *Comparaison de maturité de l'AE vs TN*

Nos résultats, représentés à la figure 5, indiquent qu'un niveau élevé d'AE ne se reflète pas forcément par une maturité significative dans la TN, en particulier pour les PME. Ceci indique qu'il faut prendre en compte des facteurs supplémentaires de mise en œuvre. Notre enquête confirme notamment une série de pratiques pour soutenir la TN aussi identifiées dans notre état de l'art : la formation au numérique, les méthodes agiles, les projets innovants transversaux, et l'accompagnement au changement. Ceux-ci sont généralement mieux maîtrisé par les GE ce qui explique leur situation dans le quadrant supérieur.

4. Discussion

Cette section complète notre analyse en la comparant à d'autres enquêtes qui explorent les thématiques portant sur les principaux défis, les artefacts de l'AE et l'évaluation de la maturité. Ces études ont été réalisées par des sociétés de conseil ou proviennent du baromètre numérique wallon.

4.1. Comparaison avec le baromètre 2022 de maturité numérique en Wallonie

L'évaluation de la contribution de l'AE a été réalisée à travers l'analyse des 4 axes du DigiScore par ordre de maturité décroissant (Digital Wallonia, 2022) :

– l'axe organisation est le plus mature (~40). Environ 50% des entreprises sont engagées dans un projet de TN contre environ 30% dans notre enquête. Les mêmes rôles sont impliqués (CEO, CIO). Des processus de formation et d'analyse de données du même ordre de grandeur sont aussi mis en place : dans environ 20% des entreprises.

– l'axe infrastructure est quasiment à égalité (~40). Les outils de communication et de collaboration sont mis en place dans plus de la moitié des entreprises, avec une

forte progression, sans doute favorisée par la crise de la COVID. Une attention à la sécurité est plus faible dans notre enquête : 11% contre 25% dans le baromètre.

– l'axe processus se situe à une maturité de 30. De nombreuses entreprises disposent déjà de processus numérisés, notamment plus de la moitié pour les services de support et 43% pour l'interaction avec le client (utilisation de CRM/ERP). Des chiffres un peu plus élevés (environ 60%) sont mentionnés dans notre enquête concernant les communications clientes et la présence d'ERP/CRM.

– l'axe stratégie est le plus faible avec une maturité de 27 et une assez grande disparité : 64% pour des grandes entreprises définissent une stratégie de TN contre 20% pour les PME. Notre enquête relevait un chiffre comparable de 25%.

Même si notre enquête était plus spécifique, les résultats obtenus sont largement alignés, confirmant ainsi la représentativité du panel collecté.

4.2. Enquête sur les défis de Deloitte

Le Digital Maturity Index Survey 2022 de Deloitte est mené environ tous les 4 ans (Deloitte, 2022) selon leur modèle de maturité (Deloitte, 2018). Nous nous sommes intéressés aux difficultés de mise en œuvre de la TN. La figure 6 illustre les difficultés en matière de sécurité des données, de gestion du changement et de complexité de la mise en œuvre. Nous y avons présenté aussi les données spécifiques à nos grandes entreprises qui sont ciblées par Deloitte. Les résultats soulignent des similitudes dans les défis rencontrés tout en mettant en évidence quelques variations : nos entreprises semblent globalement plus optimistes sur la maîtrise des obstacles.

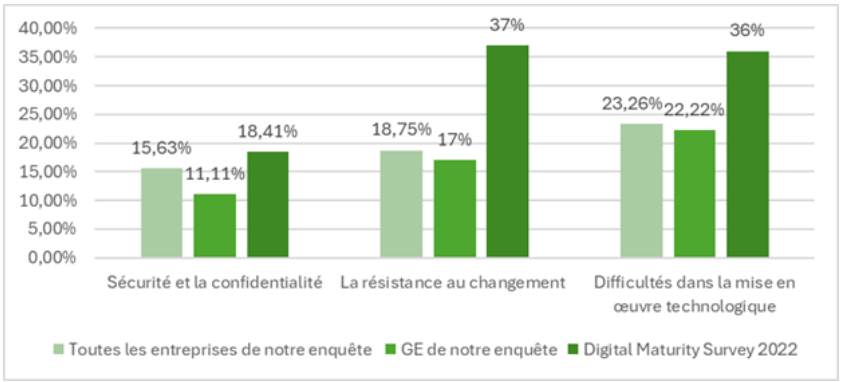


FIGURE 6. *Comparaison d'obstacles avec l'étude de Deloitte*

4.3. Comparaison d'artefacts AE mis en œuvre

Cette comparaison se concentre sur les artefacts AE en phase de planification stratégique qui a été identifiée comme lacunaire dans l'état de l'art. Nous avons comparé

les artefacts rapportés par deux enquêtes couvrant respectivement 4 entreprises du domaine financier (Grave *et al.*, 2021) et 9 entreprises du domaine de la santé (Beirnaert, 2023). La figure 7 illustre l'utilisation des analyses SWOT, ses diagrammes d'infrastructure IT, de contexte, du plan stratégique, du modèle de capacité et de processus.

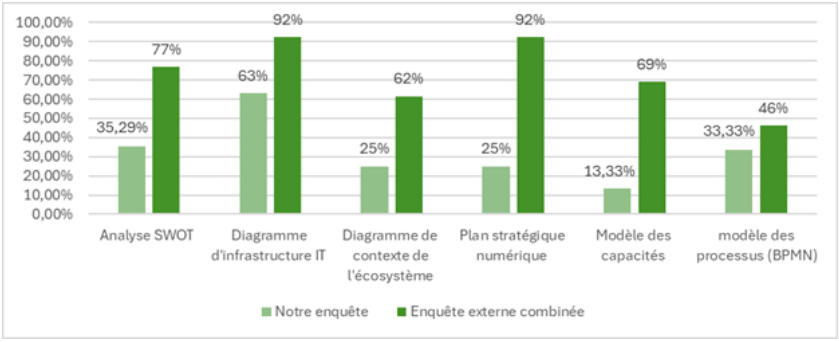


FIGURE 7. *Comparaison d'artefact AE en phase de planification stratégique*

La comparaison montre :

– un taux d'utilisation significativement plus faible dans notre enquête par rapport à la référence externe. Pour le plan stratégique numérique, une explication est que nous avons une majorité de PME peu matures sur ce point (Digital Wallonia, 2022).

– qu'en usage relatif, des points communs sont la popularité de l'analyse SWOT et des diagrammes d'infrastructure IT.

4.4. Biais possibles d'analyse

L'analyse empirique peut être affectée de plusieurs biais. Au niveau interne, les hypothèses explorent les connexions entre des facteurs qui peuvent avoir des effets plus larges notamment les techniques d'analyse retenues ou la dimension agile. Nous avons été attentif à expliciter la présence de facteurs complémentaires notamment dans l'analyse conjointe des niveaux de maturité TN et AE de l'hypothèse 3. L'influence de l'agilité n'a également pas été concluante. Au niveau externe et de la construction, nous avons déjà justifié la représentativité de l'échantillon et la pertinence des questions à la section 3.2. Au niveau statistique, la taille modeste des données récoltées a exclu des analyses détaillées par secteur. Nous nous sommes limité à la distinction la plus pertinente entre PME et GE.

5. Conclusion et perspectives

Au terme de notre analyse, on peut revenir à notre question de recherche « Comment une démarche d'architecture d'entreprise peut-elle soutenir efficacement la transformation numérique d'une organisation ? ». Notre état de l'art a clairement montré des liens significatifs entre ces deux notions avec une couverture importante de divers

aspects de l'AE sur le spectre des activités de la TN, illustrées sur la matrice du Hub Institute et l'intérêt de cet outil pour l'application ciblée de l'AE à la TN. Nous avons ensuite exploré des questions plus précises à l'aide d'une enquête menée sur un panel représentatif d'entreprises wallonnes. Les résultats ont permis de répondre partiellement à nos questions, en pointant que si l'apport de l'AE était significatif, il n'assurait pas à lui seul tous le processus de la TN. Il a aussi confirmé quelles lacunes identifiées dans l'état de l'art étaient effectivement présentes, notamment en matière de plan stratégique et d'outils pour accompagner la mise en œuvre de l'AE pour la TN. Un des obstacles majeurs, l'accompagnement au changement, est d'emblée adressé par les frameworks d'AE mais nécessite bien sûr de s'assurer de la maturité à laquelle on se situe. A cet effet, plusieurs modèles de maturité ont été explicités et la réalisation de l'enquête a aussi été l'occasion d'en expérimenter certains indicateurs.

Nos travaux futurs pourraient combler des limitations de ce travail, notamment pour disposer de données plus riches et associées à des indicateurs plus précis à la fois pour l'AE et la TN. Plutôt que de déployer une nouvelle enquête, nous envisageons de l'intégrer à la collecte du DigiScore, en ciblant un prochain baromètre du numérique et en bénéficiant de l'infrastructure de Digital Wallonia. Une piste de recherche pourrait approfondir la dynamique et les multiples dimensions des mécanismes collaboratifs d'alignement, dans le contexte numérique évoqué dans (Mendes da Silva *et al.*, 2024). L'analyse pourrait aussi élaborer des aspects laissés de côté comme les liens avec l'écosystème numérique et l'adoption de l'intelligence artificielle.

Bibliographie

Aghakhani G., Wautelet Y., Kolp M. (2021). Towards strategic support and guidance of the digital transformation: A conceptual model. In *Poem workshops.*

Avasarala V. (2020). *Ai-powered digital transformation: The ultimate guide to ai in business*. Wiley.

Beirnaert I. S. (2023). *The role of enterprise architecture in the strategic planning process: An exploratory study in the preclinical domain.* MsC, Open Universiteit.

Bouncken R., Schmitt F. (2022). SME Family Firms and Strategic Digital Transformation: Inverting Dualisms Related to Overconfidence and Centralization. *Journal of Small Business Strategy*, vol. 32, n° 3, p. 1–17.

Bounfour A. (2016). *Digital futures, digital transformation: From lean production to acceluction*. Springer.

Bughin J. *et al. (2019). Digital transformation: Improving the odds of success.* McKinsey Quarterly*, vol. 5, n° 4.*

Cantemir M. et al. (2023). Drivers of digital transformation and their impact on organizational management. *Studies in Business and Economics*, vol. 18, n° 1, p. 149–170.

CIGREF. (2008). *L'architecture d'entreprise : cadre de cohérence de la transformation du système d'information.* https://www.cigref.fr/cigref_publications/RapportsContainer/Parus2008/Cercle_Architecture_Entreprise_2008.pdf.

Deloitte. (2018). *Digital maturity model: Achieving digital maturity to drive growth.* https://tinyurl.com/deloitte-dig-mat-model.

Deloitte. (2022). *Digital maturity index survey.* https://fr.scribd.com/document/638375296/Deloitte-Digital-Maturity-Index-Survey-2022.

Digital Wallonia. (2020). *Digiscore.* https://digiscore.digitalwallonia.be.

Digital Wallonia. (2022). *Baromètre entreprises.* https://www.digitalwallonia.be/fr/publications/entreprises2022-organisation/.

Ducrey V., Vivier E. (2017). *Le guide de la transformation digitale: La méthode en 6 chantiers pour réussir votre transformation* (vol. 327). EYROLLES.

Dudézert A. (2018). La transformation digitale des entreprises. *Repères*, vol. 127.

Ebert C., Duarte C. H. C. (2018). Digital transformation. *IEEE Software*, vol. 35, n° 4, p. 16–21.

Faller C., Feldmüller D. (2015). Industry 4.0 Learning Factory for regional SMEs. *ELSEVIER*, vol. 88, n° 1, p. 88–91.

Fuchs C., Hess T. (2018). Becoming agile in the digital transformation: The process of a large-scale agile transformation. In *Icis 2018 proceedings.*

Gamache S. *et al. (2020). Evaluation of the influence parameters of industry 4.0 and their impact on the quebec manufacturing smes: first findings.* Cogent Engineering*, vol. 7, n° 1.*

Goerzig D., Bauernhansl T. (2018). Enterprise architectures for the digital transformation in small and medium-sized enterprises. Procedia CIRP*, vol. 67, p. 540–545.*

Grave F. et al. (2021). Enterprise architecture artifacts facilitating the strategy planning process for digital transformations: a systematic literature review and multiple case study. *IADIS International Journal on Computer Science and Information Systems*, vol. 16, n° 1.

Holotiuk F., Beimborn D. (2017). Critical success factors of digital business strategy. In *Proc. international conference on wirtschaftsinformatik.*

ISACA. (2012). *COBIT 5: A Business Framework for the Governance and Management of Enterprise IT.* https://www.isaca.org/COBIT/Pages/default.aspx.

ITU. (2019). *Digital Transformation and the Role of Enterprise Architecture.* https://sitic.org/digital-transformation-and-the-role-of-enterprise-architecture.

Jonathan G. M. *et al. (2023). IT Alignment: A Path Towards Digital Transformation Success.* Procedia Computer Science*, vol. 219.*

Jonathan G. M., Rusu L., Grembergen W. (2021). Business-IT Alignment and Digital Transformation: Setting a Research Agenda. Proc. of the Int. Conf. on Inf. Syst. Dev. (ISD)*.*

Jones M. D. et al. (2021). Past, present, and future barriers to digital transformation in manufacturing: A review. *Journal of Manufacturing Systems*, vol. 39, n° 6, p. 936–948.

Leipzig T. von *et al. (2017). Initialising customer-orientated digital transformation in enterprises.* Procedia Manufacturing*, vol. 8, p. 517–524.*

Mahmood F. et al. (2019). Digital organizational transformation issues, challenges and impact: A systematic literature review of a decade. *Abasyn Journal of Social Sciences*, vol. 12.

Manyika J. *et al. (2015).* Digital america: A tale of the haves and have-mores. *McKinsey.*

Mendes da Silva F. M. et al. (2024). The Impact of IT-Business Strategic Alignment on The Transformation and Operations of Pre-Digital Businesses. *Revista de Administração Contemporânea*, vol. 28.

Mergel I., al. et. (2019). Defining digital transformation: Results from expert interviews. *Government Information Quarterly*, vol. 36, nº 4.

Mhlungu N., Chen J., Alkema P. (2019). The underlying factors of a successful organisational digital transformation. *SA Journal of Information Management*, vol. 21.

OECD. (2017). *The Digital Transformation of SMEs* nº 12. OECD Report.

Office québécois de la langue française. (2007). *Grand dictionnaire terminologique.* https://vitrinelinguistique.oqlf.gouv.qc.ca/fiche-gdt/fiche/8349842/architecture-dentreprise.

Osmundsen K. *et al. (2018). Digital transformation: Drivers, success factors, and implications.* Association for Information Systems.

Ozguner Z. (2021). Evaluation of critical success factors playing roles in the digital transformation process. Journal of Economics and Business Issues*, p. 40–49. (pp. 41–42)*

Peillon S., Dubruc N. (2019). Barriers to digital servitization in French manufacturing SMEs. ELSEVIER*, vol. 146, nº 2, p. 146–150. (Pages 147-148)*

Sailer P., Stutzmann B., Kobold L. (2019, October). Successful digital transformation: How change management helps you to hold course. *Siemens IoT Services Whitepaper.*

Schallmo D., Williams C. (2017). Digital transformation of business models - best practice, enablers, and roadmap. International Journal of Innovation Management*, vol. 21, nº 8.*

SPF Economie. (2024). Digitalisation des PME. *https://economie.fgov.be/fr/themes/entreprises/pme-et-independants-en/digitalisation-des-pme.*

The Open Group. (2018). TOGAF Version 9.2. *https://www.opengroup.org/togaf*.

The Open Group. (2024). IT4IT™ Standard, Version 3.0.1. *https://www.opengroup.org/it4it.*

Triki S., Ponsard C. (2024). Enquête pme sur les pratiques ea et tn. *https://docs.google.com/forms/d/13MhvGo80yExbbfvqED9G6CfgHt0uVdZAMVhHB9BRtJQ* .

Vogelsang K. et al. (2019). Barriers to digital transformation in manufacturing: Development of a research agenda. In *Proceedings of the hawaii conference on system sciences.*

Weill P., Woerner S. L. (2013). Optimizing your digital business model. *MIT Sloan Management Review*, vol. 54, nº 3, p. 73.

Westerman G. *et al. (2012).* The digital advantage: How digital leaders outperform their peers in every industry. *Capgemini Consulting and MIT Sloan Management.*

Winer R. S., Bock G. W. (2017). Towards a taxonomy of digital business models – conceptual dimensions and empirical illustrations. Journal of Business Research*, vol. 19, p. 7-10.*

Zachman J. A. (2011). The zachman framework for enterprise architecture: Version 3.0. *https://zachman.com/framework.*

Zaoui F., Souissi N. (2020, janvier). Roadmap for digital transformation: A literature review. Procedia Computer Science*, vol. 175, p. 621–628.*

Ziyadin S. et al. (2020). Digital transformation in business. In *Proc. of the int. conference "digital transformation of the economy: Challenges, trends, new opportunities"*.

Réinternaliser le système d'information

Un système d'aide à la décision

J. Akoka[1], I. Comyn-Wattiau[2]

1. Laboratoire CEDRIC-CNAM 2. ESSEC Business School

Ce résumé reprend l'article (Akoka *& Comyn-Wattiau,* 2025) paru dans *Decision Support Systems*. Les organisations peuvent être confrontées à des problèmes lors de l'externalisation de leur système d'information. Les raisons sont notamment : un contrat défaillant, des changements organisationnels, une perte de contrôle. Dès lors, certains adoptent une solution de ré-internalisation. Compte tenu des enjeux, il est impératif de formaliser la décision. Cet article décrit un ensemble d'exigences, un cadre conceptuel, un modèle et un système d'aide à la décision.

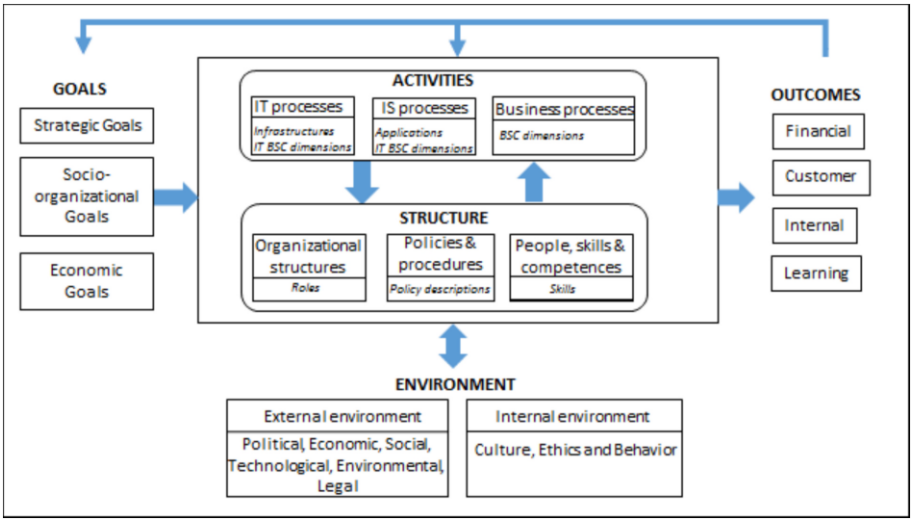


Figure 1. Le cadre conceptuel fondé sur la théorie des systèmes

L'approche comprend cinq étapes : 1) La sensibilisation conduit à l'élaboration d'un ensemble d'exigences. 2) La suggestion combine ces exigences avec la théorie des systèmes pour générer des principes de conception. 3) Le développement permet d'en déduire un cadre conceptuel (Fig. 1), ensuite traduit en un arbre de décision (Fig. 2). 4) L'application d'un ensemble de théories organisationnelles pour opérationnaliser le modèle de décision, à partir duquel un prototype est dérivé. 5) L'évaluation teste le système à l'aide d'une étude de cas, ce qui permet d'illustrer ex

post comment les décisions de l'entreprise peuvent être éclairées par plusieurs théories (institutionnelle, de l'agence, des systèmes, des coûts de transaction, etc.). Si les théories se contredisent, la pondération joue un rôle dans la hiérarchisation des recommandations. Notre système permet ainsi aux décideurs de renforcer leurs arguments et de fournir une base de discussion.

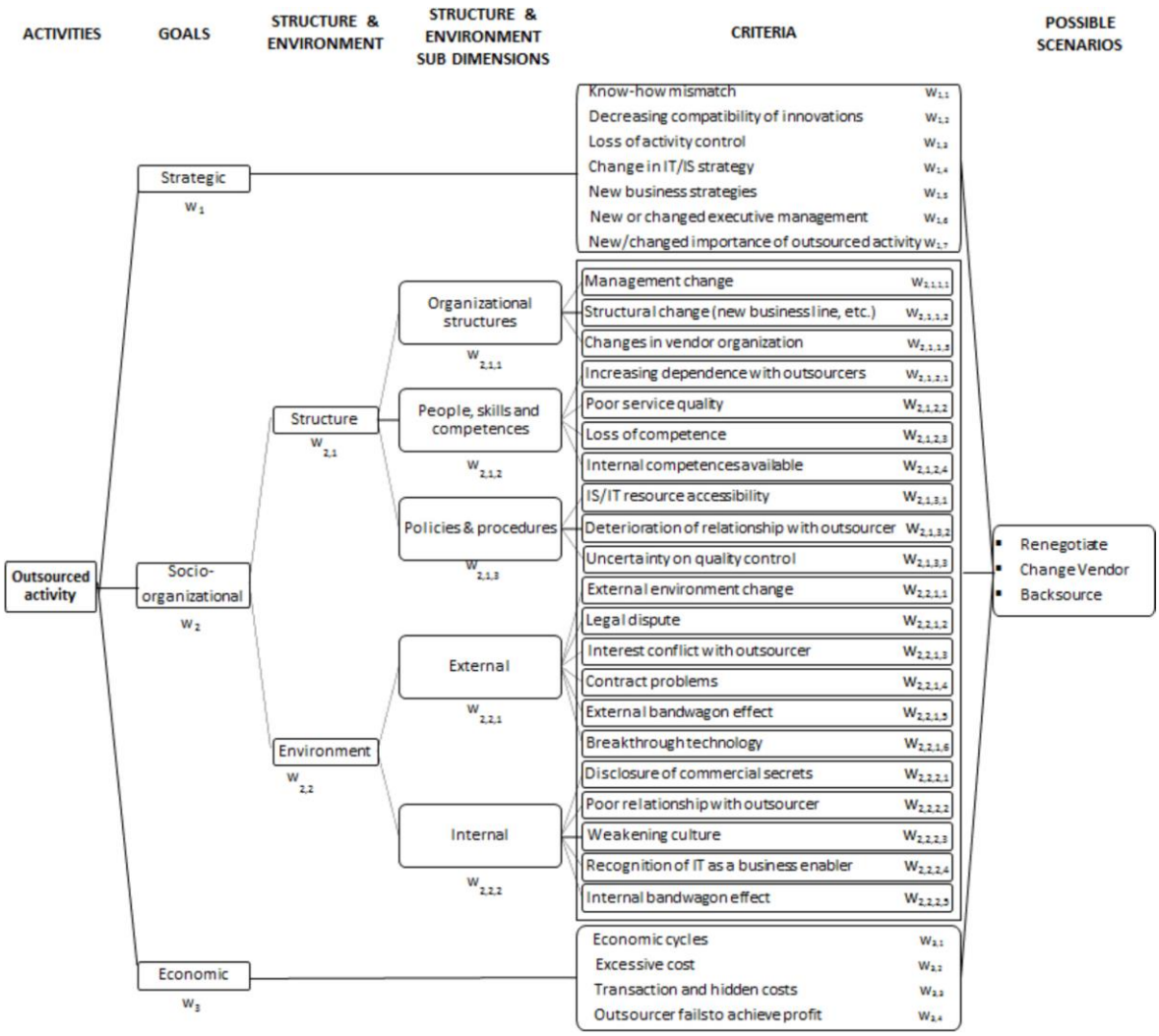


Figure 2. L'arbre de décision

Bibliographie

Jacky Akoka, Isabelle Comyn-Wattiau, IS/IT Backsourcing decision making - A design science research approach, Decision Support Systems, Volume 189, 2025, 114379, ISSN 0167-9236, https://doi.org/10.1016/ j.dss.2024.114379.

IPMD : Découverte des Modèles de Processus Intentionnel à partir des Logs d'Événements

Ramona Elali[1], Elena Kornyshova[2], Rébecca Deneckère[1], Camille Salinesi[1]

1. Paris 1 Panthéon Sorbonne, Paris, France
{ramona.elali, rebecca.deneckere, camille.salinesi}@univ-paris1.fr

2. Conservatoire National des Arts et Métiers, Paris, France
elena.kornyshova@cnam.fr

REFERENCE DE L'ARTICLE INTERNATIONAL.
Cet article est un résumé de l'article : Ramona Elali, Elena Kornyshova, Rebecca Deneckere, Camille Salinesi. IPMD: Intentional Process Model Discovery from Event Logs. In: Araújo, J., de la Vara, J.L., Santos, M.Y., Assar, S. (eds) Research Challenges in Information Science. RCIS 2024. Lecture Notes in Business Information Processing, vol 514. Springer, Cham. (2024).

1. Introduction

Les modèles de processus métier sont des représentations abstraites des activités organisationnelles, essentielles pour la gestion, l'optimisation et l'automatisation des processus. La découverte de modèles de processus à partir des logs d'événements, connue sous le nom de la fouille de processus (Aalst, 2016), est un domaine de recherche en pleine expansion. Toutefois, la majorité des approches existantes se concentrent principalement sur la perspective des séquences d'activités en négligeant souvent les intentions sous-jacentes qui guident ces activités. La prise en compte explicite des intentions permettrait d'obtenir une représentation plus riche et interprétable des processus. La fouille d'intentions est un aspect crucial pour comprendre le comportement humain. Il se concentre sur la découverte des intentions et objectifs cachés qui guident les individus dans leurs activités. Nous proposons l'approche IPMD (Intentional Process Model Discovery) qui combine la fouille des motifs fréquents (Frequent Pattern Mining), les modèles de langage de grande taille (LLM) et la fouille de processus (Process Mining) afin de construire des modèles de processus intentionnels capturant les stratégies humaines héritées de la prise de décision et de l'exécution des activités. Cette combinaison vise à identifier des séquences récurrentes d'actions révélant les stratégies (motifs

récurrents d'activités) que les utilisateurs appliquent couramment pour atteindre leurs intentions. Ces motifs sont ensuite utilisés pour construire un modèle de processus intentionnel basé sur le formalisme MAP (Rolland *et al.*, 1999) fondé sur la découverte de stratégies.

2. Proposition

Notre approche IPMD, qui propose la découverte des modèles de processus intentionnels, vise à révéler les intentions et stratégies des utilisateurs derrière leurs activités en utilisant des logs d'activités. Contrairement aux modèles de processus traditionnels axés sur les activités, qui décrivent ce que l'utilisateur fait et quand, cette approche met l'accent sur le pourquoi des comportements des utilisateurs, offrant ainsi une vision plus abstraite et orientée vers les objectifs. L'approche IPMD adopte une méthodologie ascendante (bottom-up), où le modèle de processus est construit automatiquement à partir des logs d'activités, sans nécessiter de connaissances préalables sur les objectifs de l'utilisateur. Elle repose sur trois niveaux hiérarchiques : le niveau opérationnel qui représentent les activités individuelles, le niveau stratégique qui identifient les motifs d'activités regroupés en stratégies, et le niveau intentionnel qui illustrent les intentions ou objectifs globaux derrière ces stratégies. Pour atteindre cet objectif, IPMD combine trois techniques principales : la fouille des motifs fréquents pour découvrir des motifs de stratégie à partir des logs d'activités, la fouille de processus pour construire des modèles de processus basés sur ces motifs, et le modèle de langage de grande taille pour nommer et décrire les stratégies et intentions découvertes.

Une contribution importante de ce travail est l'extension du méta-modèle MAP en ajoutant les concepts d'activité et de motif de stratégie. Cette extension permet d'adapter le méta-modèle à une approche ascendante, contrairement aux approches descendantes (top-down) nécessitant des interviews ou des modèles existants. L'approche a été appliquée au log de données qui décrit les habitudes quotidiennes d'un individu à domicile. En reliant automatiquement les activités aux intentions des utilisateurs à travers les motifs de stratégie, cette méthode améliore la qualité des modèles de processus, qui va faciliter dans le futur la personnalisation des recommandations et rendre la découverte des processus plus efficace et scalable.

Bibliographie

Aalst W.M.P. (2016). *Process Mining: Data Science in Action*. Springer, Heidelberg ISBN : 978-3-662-49850-7.

Rolland, C., Prakash, N., and Benjamen, A.: A Multi-Model View of Process Modelling. Requirements Engineering, pages 169 – 187. (1999).

EM-BPMN+X : Une Méthode Générique d'Aide à la Mise en Oeuvre des Extensions de BPMN Valides

Mariam Ben Hassen [1], Faïez Gargouri [2]

1. University of Sfax, ISIMS, MIRACL Laboratory - B.P. 242, 3021 Sfax, Tunisia
University of Gafsa, Higher Institute of Business Administration, 2112 Gafsa, Tunisia
mariem.benhassen@isims.usf.tn

2. University of Sfax, ISIMS, MIRACL Laboratory - B.P. 242, 3021 Sfax, Tunisia
faiez.gargouri@isims.usf.tn

RÉSUMÉ. La gestion des enjeux d'intégrité, de flexibilité et d'interopérabilité dans les systèmes d'information d'entreprise (EIS) est souvent freinée par les problèmes dits des « 3-Fit » : vertical, horizontal et transversal. Pour y remédier, cette recherche propose des solutions visant à structurer et enrichir la Vue Métier des EIS, en s'appuyant sur des approches avancées de modélisation des processus métier (BP). Bien que BPMN 2.0.2 soit un standard largement adopté pour modéliser et exécuter les workflows organisationnels, il reste peu adapté aux processus complexes, flexibles, hautement dynamiques, interactifs et à forte intensité de connaissances. Pour combler ces limites, nous introduisons EM-BPMN+X, une méthodologie rigoureuse pour développer des extensions valides de BPMN 2.0.2, spécifiques à des domaines, et fondées sur des ontologies noyaux de domaine. Cette approche fournit un cadre holistique articulant l'analyse du domaine d'extension, la modélisation conceptuelle et la mise en œuvre de méta-modèles, afin de renforcer la clarté, la cohérence et la réutilisabilité des modèles de BP.

ABSTRACT. The challenge of ensuring integrity, flexibility, and interoperability in Enterprise Information Systems (EISs) is hindered by the "three-fit" barrier, which encompasses vertical, horizontal, and transversal fit problems. This paper addresses these challenges by refining the Business View of EIS through the integration of advanced business process modeling (BPM) approaches. Among these, Business Process Model and Notation (BPMN 2.0.2) is widely recognized for structuring, executing, and analyzing organizational workflows. While BPMN enhances stakeholder communication and operational efficiency, it lacks the expressive power to model complex, flexible, highly dynamic, interactive, and knowledge-intensive processes. To bridge this gap, we introduce EM-BPMN+X (Methodology for the Development of BPMN Plus Extensions), a structured and rigorous approach for developing valid domain-specific BPMN 2.0.2 extensions grounded in core domain ontologies. This methodology provides a clear framework for linking domain analysis, conceptual modeling, and meta-model implementation, strengthening clarity, consistency, and reusability in BPM.

Mots-clés : Systèmes d'information d'entreprise, modélisation des processus métier, gestion des connaissances, processus métier sensibles, ontologies noyaux de domaine, BPMN 2.0.2, mécanisme d'extension, méthodologie de recherche en science de la conception.

KEYWORDS : Enterprise Information Systems, Business Process modeling, Knowledge Management, Sensitive Business Process, Core Domain Ontologies, BPMN 2.0.2, Extension Mechanism, Design Science Research Methodology.

1. Problématique et Motivation

Les entreprises contemporaines évoluent dans des environnements dans des environnements dynamiques, complexes et hautement concurrentiels. Pour renforcer leur performance, elles doivent adopter une approche centrée sur les processus métier (Business Process – BP),

favorisant l'agilité, l'optimisation des chaînes de valeur et l'efficacité opérationnelle. Leur compétitivité repose ainsi sur trois piliers essentiels : l'intégrité des BP, l'interopérabilité avec les parties prenantes externes, et la capacité d'adaptation aux évolutions du marché (Ben Hassen *et al.*, 2024a). Bien que les Systèmes d'Information d'Entreprise (EIS) soient conçus pour répondre à ces enjeux, ils rencontrent souvent des difficultés dues à l'hétérogénéité organisationnelle et à la rigidité de leurs architectures.

Pour y remédier, notre recherche s'appuie sur l'approche d'urbanisation (Fournier-Morel *et al.*, 2008), qui structure le développement des EIS selon quatre vues complémentaires (Ben Hassen *et al.*, 2024a) : (1) la *Vue Métier*, qui définit les événements et organise les activités métier; (2) la *Vue Fonctionnelle*, qui identifie les rôles, les connaissances et les objets métier ; (3) la *Vue Applicative*, qui spécifie les solutions logicielles nécessaires à l'exécution des processus ; et (4) la *Vue Physique*, qui couvre l'infrastructure technique (voir Figure 1). Cette architecture multicouche permet d'aligner efficacement les objectifs stratégiques, les processus opérationnels, les outils technologiques et les infrastructures physiques. Toutefois, il est souvent confronté aux problèmes dits des « trois-fit » (Figure 1) : (1) le « Fit vertical », résultant d'un désalignement entre les couches métier et technique, affectant l'intégrité et l'extensibilité du système ; (2) le « Fit horizontal », lié à une faible interopérabilité interne et une rigidité organisationnelle, limitant la flexibilité ; (3) le « Fit transversal », résultant d'un manque d'ouverture limitant l'interopérabilité externe. Ces obstacles découlent principalement d'un manque de spécifications formelles garantissant la cohérence, l'adaptabilité et l'intégration fluide des dimensions métier et technique. Leur dépassement exige la conception d'EIS intrinsèquement flexibles, ouverts et interopérables.

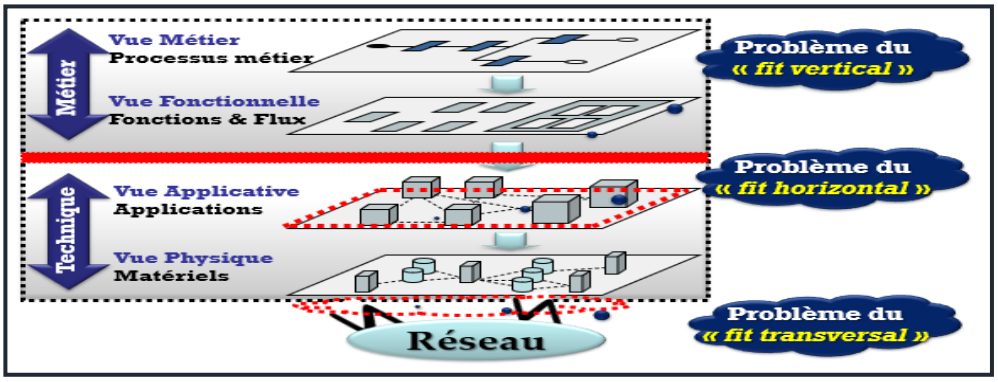


Figure 1. Modèle de référence de l'architecture des SIE : Problèmes de « 3Fit » (Ben Hassen et al., 2024a)

Dans un contexte organisationnel de plus en plus axé sur la collaboration, l'agilité et la gestion des connaissances, les entreprises prennent conscience de la valeur stratégique des savoirs mobilisés au sein de leurs BP. L'identification, la formalisation, le partage et la réutilisation de ces connaissances — qu'elles soient tacites ou explicites, individuelles ou collectives — deviennent des leviers essentiels pour améliorer la performance et préserver un avantage concurrentiel durable. Les BP sensibles à forte intensité de connaissances englobent des activités critiques s'appuyant sur l'acquisition, la diffusion, le partage, la création et la réutilisation intensive des connaissances cruciales. Leur exécution requiert une collaboration étroite entre acteurs intra- et inter-organisationnels, impliquant transfert, création et application de connaissances pour atteindre des résultats à forte valeur ajoutée. Leur nature souvent partiellement ou non structurée, interactive, évolutive, et complexe rendent leur modélisation, leur exécution et leur pilotage particulièrement difficiles et critique (Ben Hassen *et al.*, 2024b).

Pour surmonter les problèmes des « trois-fit », cette recherche propose d'améliorer la *Vue Métier* des EIS par l'intégration d'approches solides de modélisation des BP (BPM). Dans le

domaine de *Business Process Management*, les modèles de BPs sont devenus des outils incontournables pour concevoir, executer, analyser et piloter les workflows organisationnels. Ils facilitent la compréhension et la communication entre parties prenantes, soutiennent la transformation numérique et favorisent l'amélioration continue. L'expressivité des modèles contemporains de BP repose sur six dimensions complémentaires (Ben Hassen *et al.*, 2024b) : (1) La *dimension fonctionnelle* : elle couvre les activités critiques à forte intensité de connaissances cruciales, les actions collectives, la collaboration, les conversions de connaissance, etc. (2) La *dimension connaissance* : elle distingue les données, informations et connaissances, catégorise les types de connaissances (individuels/collectifs, tacites/explicites, factuelles/procédurales, etc.), identifie les sources et différencie les flux d'information de ceux de connaissance. (3) La *dimension organisationnelle* : elle considère les entités agentives impliquées dans la réalisation des BP et leurs interactions axées sur le partage et la création des connaissances. (4) La *dimension informationnelle* : elle décrit les différents types d'information, leurs sources, et la dynamique des flux d'information au sein des activités. (5) La *dimension comportementale* : elle modélise les dynamiques de transfert et de conversion des connaissances et informations. (6) La *dimension intentionnelle* : elle intègre les informations contextuelles et les intentions (distales) guidant la planification, l'exécution et le contrôle des actions vers des objectifs stratégiques. Pour enrichir la modélisation multi-dimensionnelle des BP, un langage approprié de BPM doit intégrer explicitement ces six dimensions issues de l'articulation entre BPM et Knowledge Management (KM). Aujourd'hui, BPMN 2.0.2 (OMG, 2013) est reconnu comme le standard de référence en BPM grâce à sa richesse sémantique, sa clarté, sa large adoption, la clarté de son méta-modèle, la disponibilité d'outils de modélisation, ainsi que sa flexibilité (Ben Hassen *et al.*, 2019). Toutefois, ses concepts restent trop génériques pour couvrir de manière exhaustive tous les aspects pertinents de modélisation des BP complexes, dynamiques et à forte intensité de connaissance. Afin de pallier les limitations actuelles de BPMN 2.0.2 et répondre efficacement aux problèmes des « trois-fit », cette recherche propose une méthodologie générique, rigoureuse et holistique pour concevoir des extensions spécifiques du langage BPMN. Ces extensions sont à la fois compréhensibles et pleinement conformes à sa spécification officielle. L'approche proposée, nommée **EM-BPMN+X** (*Methodology for the Development of BPMN Plus Extensions*), s'appuie sur des travaux de référence éprouvés (Stroppi *et al.*, 2011 ; Braun *et al.*, 2014 ; 2015 ; 2016 ; Ben Hassen *et al.*, 2017a ; 2017b ; 2022 ; 2024a ; 2024b ; Ben Hassen and Gargouri, 2024) et fournit un cadre structuré pour guider le processus de développement d'extensions adaptées à des domaines spécifiques, notamment ceux à forte intensité de connaissances. La méthode se décline en trois phases : (1) *Analyse et conceptualisation du domaine cible*, pour identifier les concepts clés et les exigences spécifiques (dans notre cas, la modélisation des BP sensibles (SBP) dans une perspective de KM); (2) *Préparation de l'extension*, en sélectionnant les éléments à intégrer et en définissant leur correspondance avec la méta-structure BPMN ; (3) *Définition des syntaxes abstraite et concrète*, conformément au mécanisme d'extension de BPMN 2.0.2, assurant l'interopérabilité avec les outils existants. EM-BPMN+X garantit ainsi des extensions robustes, cohérentes et alignées avec le standard BPMN, facilement adaptables à divers domaines. En comblant l'écart entre l'analyse du domaine, la conceptualisation et l'implémentation du métamodèle étendu, elle renforce la clarté, la réutilisabilité et la cohérence des langages BPM spécifiques.

L'objectif visé est de permettre à BPMN de s'adapter efficacement à des contextes organisationnels complexes, dynamiques et à forte intensité d'informations et de connaissances, en comblant les insuffisances actuelles des outils de modélisation standards. Le développement de cette approche s'inscrit dans le cadre méthodologique rigoureux de la Design Science Research Methodology (DSRM) (Hevner and Chatterjee, 2010 ; Peffers *et al.*, 2018), assurant à la fois la rigueur scientifique et la pertinence pratique des artefacts conçus. Ce travail s'intègre dans un projet de recherche à long terme portant sur la spécification des BP sensibles dans une perspective de KM, fondé sur des exigences réelles et structuré en cycles itératifs de conception, d'évaluation et de validation expérimentale (Ben Hassen *et al.*, 2017a ; 2017b ; 2022 ; 2024a ; 2024b ; Ben Hassen and Gargouri, 2024). Il vise à fournir une solution innovante, flexible et évolutive pour la modélisation et la gestion des BP, notamment dans des domaines à haute

complexité tels que la santé, la finance, la biotechnologie, la cybersécurité, la gestion de crise, les télécommunications ou encore l'éducation. En proposant une approche rigoureuse, évolutive et conforme aux standards, EM-BPMN+X constitue un cadre méthodologique structurant et prometteur, permettant d'améliorer la modélisation, l'exécution et la performance des BP. Elle contribue ainsi à renforcer l'efficacité organisationnelle, la prise de décision stratégique et la maîtrise des risques, quels que soient les contextes d'application. Conformément aux phases de la méthode DSRM (Hevner and Chatterjee, 2010 ; Peffers *et al.*, 2018), le reste de l'article est structuré comme suit : introduit la problématique et les motivations de notre travail. Elle commence par un rappel des principes d'extensibilité de BPMN, suivi d'un aperçu des principales méthodes valides de conception d'extensions reconnues dans la littérature, notamment celles de Stroppi *et al*. et de Braun *et al*. Elle se poursuit par une revue systématique de la littérature, permettant d'établir l'état de l'art des extensions BPMN et d'en identifier les lacunes encore présentes dans ce champ de recherche. La section 3 définit nos objectifs de recherche et les solutions proposées face aux défis actuels du BPM. La section 4 présente la conception et le développement de l'artefact, à savoir la méthodologie EM-BPMN+X. La section 5 présente l'évaluation de notre contribution. Enfin, la section 6 conclut l'article en récapitulant les apports majeurs et en ouvrant sur des perspectives de recherche futures.

2. Extensibilité de la spécification BPMN : État de l'art

2.1. Mécanisme d'extension de BPMN 2.0.2 et ses lacunes

Dans l'optique d'étendre les cas d'utilisation de la spécification BPMN 2.0.2 à l'aide de concepts spécifiques à divers domaines, l'OMG (2013) a introduit un mécanisme d'extension par addition. Ce mécanisme permet d'ajouter des concepts et attributs propres à un domaine d'application, tout en garantissant la validité des éléments de base du langage BPMN. Reposant sur l'architecture MOF (OMG, 2014), ce mécanisme s'appuie sur des éléments d'extension génériques tels que : `ExtensionDefinition`, `ExtensionAttributeDefinition`, `ExtensionAttributeValue` et `Extension`.

Bien que ce mécanisme fournisse des structures et relations pertinentes pour l'extension du métamodèle BPMN, il présente néanmoins des ambiguïtés syntaxiques et lacunes formelles susceptibles de générer des incertitudes lors de sa mise en œuvre. Par exemple, la spécification ne définit pas clairement les types des nouveaux attributs à intégrer. Il demeure également flou si une extension BPMN doit être considérée comme une nouvelle version du métamodèle ou comme un profil, au sens des profils UML. De plus, la spécification ne fournit aucune directive méthodologique encadrant l'utilisation appropriée des éléments d'extension standard, ni pour le développement cohérent d'extensions, ni pour la définition de notations graphiques adaptées. En l'absence de telles recommandations, la compréhensibilité, la qualité et la capacité d'échange des modèles BPMN étendus se trouvent fortement compromises (Ben Hassen *et al.*, 2017a ; 2017b).

2.2. Fondements méthodologiques et premiers prolongements pour l'extension systématique de BPMN

À ce jour, peu de travaux ont traité en profondeur le développement méthodologique des extensions BPMN selon le mécanisme d'extension proposé par l'OMG. L'étude pionnière en la matière est celle de Stroppi *et al.* (2011), qui proposent une méthode fondée sur l'approche MDA, encadrant la création d'extensions valides depuis leur conceptualisation jusqu'à leur sérialisation. Elle se compose de trois phases principales : (1) la définition d'un modèle conceptuel de domaine de l'extension (CDME), distinguant les concepts standards de BPMN et ceux spécifiques à l'extension ; (2) la transformation du CDME en un modèle BPMN+X valide, à l'aide de profils UML, stéréotypes, et règles de transformation de modèles ; (3) la génération d'un schéma XML permettant la sérialisation du modèle et la création de workflows

exécutables. Cette méthode fournit une base structurée pour définir la syntaxe abstraite des extensions BPMN. Toutefois, elle présente certaines limites : les premières phases manquent de guidage, notamment pour l'analyse des exigences du domaine et la préparation du CDME. De plus, l'absence d'évaluation sémantique approfondie entre concepts du domaine et ceux de BPMN peut entraîner des redondances ou une sous-exploitation des éléments existants. Ces limites nécessitent une analyse préalable plus poussée de la spécification BPMN afin de garantir la pertinence des extensions.

Les travaux de Braun et al. (Braun et al., 2015 ; 2016)) viennent enrichir ceux de Stroppi et al. pour mieux couvrir l'analyse du domaine et renforcer la conformité avec BPMN. Leur démarche, appliquée au domaine de l'E-Santé pour modéliser les processus cliniques (clinical pathways), s'articule en six étapes : (1) analyse des exigences du domaine, (2) justification de l'adéquation de BPMN et construction d'une ontologie de domaine des processus cliniques, (3) vérification de l'équivalence sémantique entre les concepts du domaine et ceux de BPMN, (4) définition du modèle CDME basé sur l'approche de Stroppi *et al.*, (5) génération d'un modèle BPMN+X valide et (6) élaboration de la syntaxe concrète via des notations graphiques. Malgré ces améliorations, plusieurs limites subsistent. L'analyse des exigences et la vérification sémantique demeurent superficielles, et le méta-modèle étendu manque de précision. L'ontologie proposée reste peu formalisée, avec des concepts flous sur le plan sémantique. Enfin, l'approche n'est pas implémentée, ce qui limite son opérationnalisation.

2.3. Une revue systématique de la littérature sur les extensions de BPMN

Comme mentionné précédemment, il n'existe pas de méthode scientifique détaillée pour appliquer les différents éléments d'extension définis par le méta-modèle de BPMN. Ainsi, le développement d'extensions BPMN reste largement « ad hoc », ce qui est insatisfaisant, notamment au regard de l'approche DSRM (Hevner and Chatterjee, 2010). En particulier, on note l'absence de directives pour la conceptualisation du domaine ainsi qu'un manque d'analyse sémantique approfondie entre les concepts du domaine et les éléments BPMN. Les approches précédemment évoquées tentent d'y répondre, mais présentent des lacunes : absence d'analyse de domaine (Stroppi *et al*., 2011), manque de rigueur conceptuelle ou faible niveau de détail et d'applicabilité (Braun *et al*., 2015). Dans ce contexte, le développement d'une démarche holistique pour des extensions BPMN valides s'avère essentiel. Elle doit garantir à la fois la conformité au standard, la clarté des modèles et une meilleure spécification graphique des SBP dans une perspective de gestion des connaissances cruciales. C'est pourquoi nous avons mené une revue bibliographique systématique et descriptive des travaux récents les plus représentatifs de l'état de l'art en matière d'extension de BPMN (versions 2.0 et 2.0.2).

2.3.1. Framework d'analyse des extensions de BPMN

À l'issue d'un processus rigoureux de collecte et de filtrage de la littérature (environ 80 publications parues entre 2011 et 2024), 60 propositions d'extension de BPMN ont été sélectionnées pour une analyse approfondie et comparative, selon un ensemble de critères définis. Cette analyse poursuit plusieurs objectifs : (i) Dresser un état de l'art exhaustif sur les extensions de BPMN, en identifiant les tendances d'évolution au cours de la dernière décennie ; (ii) Évaluer la conformité des extensions publiées à la suite de la spécification officielle du mécanisme d'extension BPMN par l'OMG ; (iii) Examiner les caractéristiques des extensions du point de vue de leur domaine d'application et de leur processus de conception méthodologique. Cela inclut notamment les questions suivantes : quels sont les secteurs ciblés et les objectifs visés ? quels formats de représentation sont adoptés ? les extensions proposées respectent-elles les exigences syntaxiques et sémantiques du standard BPMN ? comment sont-elles démontrées, implémentées et évaluées ? ;
(iv) Enfin, mettre en évidence les principales lacunes et axes d'amélioration encore présents dans ce champ de recherche.

Nous avons distingué deux catégories de classification des extensions de BPMN selon leur finalité. La première catégorie « BPM spécifique à un domaine » concerne les extensions destinées à représenter ou à gérer les BP d'un domaine particulier, dont nous citons par exemple : l'E-santé et la modélisation des processus cliniques (e.g., Barun *et al.*, 2015; 2016), (Onggo *et al.*, 2018), (Neumann *et al.*, 2019), (Pufahl *et al.*, 2022), (Szelągowski *et al.*, 2022)), le système cyber-physique (Graja et al., 2017), la gestion des interventions en cas de catastrophe (Betke et Seifert, 2017), le cybersécurité (Chergui et Benslimane, 2020), l'externalisation des BP vers le Cloud (e.g., (Louar *et al.*, 2018), le Cloud computing (e.g., (Dukaric and Juric, 2018), (Zarour *et al.*, 2019), l'IoT (e.g., (Vogel *et al.*, 2018)), le manufacturing (e.g., (Polderdijk *et al.*, 2018), (Abouzid et Saidi, 2019)), etc. Tandis que la deuxième catégorie « amélioration de BPM » englobe soit les extensions qui visent à améliorer le langage BPMN (e.g., expressivité (e.g., (Stroppi *et al.*, 2011 ; 2012), (Braun and Esswein, 2014), (Polančič, 2020)), complexité (e.g., (Onggo *et al.*, 2018), (Carvalho *et al.*, 2018), (Szelągowski *et al.*, 2022), (Strutzenberger *et al.*, 2024), (Skouti *et al.*, 2024)), variabilité et flexibilité (e.g., (Ben Said *et al.*, 2018)), soit les extensions qui spécifient les exigences de BPM en termes de différents critères (e.g., performance, coût, sécurité, conformité et qualité (e.g., (Heguy *et al.*, 2019), soit les extensions qui contribuent aux activités de gestion des BP (e.g., simulation (e.g., (Cartelli *et al.*, 2016), exécution (e.g., (Neumann et al., 2019), (Strutzenberger *et al.*, 2024)), monitoring/surveillance (Ramos-Merino *et al.*, 2019), process mining (Strutzenberger *et al.*, 2024)). Ces différentes extensions sont indépendantes d'un domaine spécifique (i.e., qui peuvent être utilisées dans n'importe quel domaine).

Dans l'Annexe de ce papier (voir note de bas de page)[1], nous présentons un extrait de notre framework d'analyse et comparaison des principales extensions de BPMN. Nous avons évalué et comparé l'ensemble des extensions de BPMN identifiées selon un ensemble de dimensions et critères : Attributs de base de l'extension (auteur et année de la publication, titre de l'extension, domaine ciblé et objectif principal de l'extension (Descriptive, Analytique ou Exécution)) ; Conformité au mécanisme d'extension standard de BPMN (type de définition de l'extension (la façon dont l'extension est définie et expliqué), syntaxe abstrait de l'extension (définition d'un méta-modèle), syntaxe concrète de l'extension (définition de nouvelles notations), conflits sémantiques avec le standard BPMN) ; Méthode d'extension appliquée/ les aspects méthodologiques et d'analyse de domaine (analyse des exigences, vérification de la correspondance sémantique des concepts de domaine avec les éléments BPMN, réutilisation des artefacts de domaine existants (ontologies, profils UML, concepts de modélisation de domaine, etc.), modèle de processus/ approche méthodologique appliquée) ; Extension spécifique à un domaine (toutes les extensions et les adaptations/personnalisations pour l'intégration des aspects spécifiques à un domaine dans BPMN (i.e., les nouvelles éléments, relations, propriétés et diagrammes ajoutés ainsi celles qui sont étendus ou adaptés/spécifiés et les styles de l'extension), les aspects d'implémentation et d'évaluation.

2.3.2. Synthèse de l'analyse bibliographique

Notre revue systématique des extensions existantes de BPMN révèle différentes constatations et limitations :

– **Spécification et validation limitées du méta-modèle** : Il existe peu d'extensions (moins de la moitié (33,33%)) qui expliquent le méta-modèle de l'extension envisagée et considèrent sa validité en termes de méta-modèle de BPMN étendu conçus conformément au mécanisme d'extension et aux préconisations spécifiées par l'OMG (OMG, 2013). Cela entrave la compréhensibilité des extensions développées (par les adopteurs de BPMN) et empêche l'interchangeabilité des modèles. Rappelons qu'une extension est considérée comme

[1]https://github.com/mariem08790390/SBP-BPMNX-Supplementary-Material/blob/main/Annexes%201%262_%20Classification%20des%20extensions%20de%20BPMN%20dans%20la%20litt%C3%A9rature.pdf

conforme/valide si elle est représentée par un méta-modèle d'extension de BPMN ou un XML Schema défini dans la documentation officielle du langage BPMN (en termes d'utilisation du mécanisme d'extension du méta-modèle de BPMN 2.0). Parmi les travaux qui proposent des extensions valides, nous citons ceux de : (Stroppi *et al.,* 2012), (Braun and Esswein, 2015), (Braun *et al.,* 2015 ; 2016), (Ben Said *et al.,* 2018), (Betke et Seifert, 2017), (Heguy *et al.,* 2019), (Neumann *et al.,* 2019), (Zarour *et al.,* 2019), (Chergui and Benslimane, 2020), (Skouti et al., 2024) (la plupart de ces extensions sont basées sur l'approche de Stroppi *et al.* (2011)). Évidemment, ce faible taux de conformité est dû, notamment, aux aspects syntaxiques du mécanisme d'extension et à l'absence d'une méthodologie standard et détaillée pour la mise en œuvre de ce mécanisme et des extensions envisagées dans la spécification du BPMN puisque l'OMG ne spécifie que les formats de représentation des extensions (*i.e.,* le méta-modèle MOF et le Schéma XML). Par ailleurs, compte tenu du nombre important des méta-modèles définis dans la documentation de BPMN, les auteurs ont du mal à trouver lequel doit être étendu. D'autres causes sont également possibles, comme la longueur de la documentation BPMN (plus de 500 pages) et le manque de clarté sur la manière d'étendre les méta-modèles.

- **Non-conformité avec le méta-modèle BPMN** : Seulement 28,33 % des extensions sont définies et spécifiées comme des méta-modèles de BPMN étendus en appliquant le mécanisme d'extension de BPMN. Les autres extensions ne sont pas conformes au méta-modèle de BPMN. Ces dernières sont définies soit par une approche de méta-modélisation dédiée (« Propre Ext ») (45%) utilisant UML ou des expressions OCL (*e.g.,* (Dukaric and Juric, 2018), (Ramos-Merino *et al.,* 2019), (Polančič, 2020)). Soit ces extensions ne sont pas conçues sous forme de méta-modèles et sont définies que graphiquement par de nouveaux éléments de notation (*e.g.,* par de nouvelles icônes) (« Notation Ext Propre ») (11,67%) (*e.g.,* (Ammann, 2012)), (Carvalho *et al.,* 2018), (Santra and Choudhury, 2018). 15 % des extensions ne présentent aucune définition de l'extension (*e.g.,* (Supulniece *et al.,* 2010), (Polderdijk *et al.,* 2018)). En somme, la majorité des extensions ne sont pas conformes à la norme BPMN et les résultats fournis ne sont pas compréhensibles en termes de réutilisation.
- **Focus sur la syntaxe concrète :** Le développement de l'extension BPMN est actuellement axé sur la syntaxe concrète. 60% des extensions analysées présentent une syntaxe concrète étendue décrivant explicitement de nouvelles notations graphiques. 28,33 % des extensions présentent implicitement de nouveaux éléments graphiques démontrés par des exemples concrets dans des modèles de BP. Les autres extensions (11,67%) ne définissent ou n'expliquent aucun type d'extension graphique.
- **Compatibilité sémantique avec BPMN :** La majorité des extensions ne contiennent pas de conflits sémantiques avec le standard BPMN. La spécification BPMN prétend ne pas contredire la sémantique d'aucun élément.
- **Manque de rigueur méthodologique dans la conception :** Moins de priorité est accordée à la présentation méthodique claire du processus de conception. 36,67% n'appliquent aucune démarche méthodologique. Ces extensions sont développées de manière ad hoc, ce qui empêche l'évaluation de la reproductibilité et de la compréhensibilité. 31,67 % des extensions ont été conçues sur la base du modèle d'extension de BPMN (19 au total), où douze sont conçues et développées méthodiquement par application de la méthode de conception des extensions de BPMN de Stroppi *et al.* (2011) (*e.g.,* (Braun *et al.,* 2015), (Betke and Seifert, 2017), (Vogel et al., 2018), (Neumann *et al.,* 2019), (Chergui et Benslimane, 2020), (Polančič, 2020)). Soulignons que les extensions de (Braun *et al.,* 2015) et (Neumann *et al.,* 2019) sont conçues par application de la méthode d'extension de BPMN intégré de (Braun and Schlieter, 2014) (qui est basée sur l'approche de Stroppi et étendu en ce qui concerne l'analyse de domaine et sa conceptualisation). Rappelons que la méthode de Stroppi et al. (2011) et la méthode de Braun and Schlieter (2014) sont conformes/valide puisqu'elles sont basées sur le méta-modèle MOF en spécifiant les correspondances avec ses méta-classes à travers des stéréotypes. 31,67 % d'extensions (19 extensions) ont été conçues sur la base de procédures décrites individuellement (*e.g.,* (Dukaric and Juric, 2018), (Santra and Choudhury, 2018), (Abouzid and Saidi, 2019), (Ramos-Merino *et al.,* 2019), (Zarour *et*

al., 2019)). Concernant le critère de l'*analyse des exigences*, environ 78,33% des extensions analysées fournit des exigences à l'extension. La moitié est explicitement mentionnée (*e.g.*, par un ensemble d'exigences E1 à En (Braun *et al.*, 2014). Le reste des travaux décrit les exigences implicitement dans l'introduction ou à travers une description du contexte d'application (*e.g.*, (Ben Said *et al.*, 2018), (Heguy *et al.*, 2019)). 80 % des publications ont conçu l'extension spécifique sans examen approfondi de la question de savoir si chaque exigence d'extension nécessite nécessairement un nouveau concept d'extension. Il n'y a aucune comparaison de correspondance entre la sémantique des éléments de domaine et des éléments standards de BPMN. Nous supposons que plusieurs extensions BPMN n'exploitent pas l'intégralité de l'expressivité de BPMN. 20 % ont mené une discussion pour chaque concept (*e.g.*, (Braun *et al.*, 2015), (Graja *et al.*, 2017), (Neumann *et al.*, 2019), (Polančič, 2020). En outre, 65,45% des extensions de BPMN utilisent/intègrent des artefacts spécifiques à un domaine. Par exemple, les profils UML (Pillat *et al.*, 2015), (Bocciarelli *et al.*, 2016), des concepts de modélisation de domaine (Supulniece *et al.*, 2010), ((Braun and Schlieter, 2014), (Ben Said *et al.* 2018), des ontologies (Braun *et al.*, 2015 ; 2016), (Chergui and Benslimane, 2020) ou des exigences (Graja *et al.*, 2017), (Louar *et al.*, 2018), (Polančič, 2020 sont réutilisés dans la conception de l'extension.

- **Évaluation et mise en œuvre limitées :** Les extensions de BPMN sont rarement évaluées et peu implémentées malgré l'existence de plusieurs outils de modélisation extensibles.
- **Extensions BPMN pour KM (dans les BP à forte intensité de connaissances) :** Parmi les travaux analysés, très peu d'initiatives intègrent explicitement les aspects de KM dans les modèles étendus de BPMN, notamment ceux de Supulniece *et al.* (2010) et Ammann (2012). Ces propositions restent limitées, ne couvrant que partiellement certains éléments du domaine KM (comme les connaissances tacites ou les modes de conversion des connaissances), sans offrir une modélisation complète et cohérente de la dimension « connaissance » dans les BP à forte intensité de connaissances (comme les SBP (Ben Hassen *et al.*, 2024b). Les modèles représentés n'intègrent pas parfaitement et complètement les aspects pertinents de la dimension connaissance (e.g., modélisation des connaissances mobilisées et produites par les activités organisationnelles, distinction entre les flux de connaissance et les flux d'information, modélisation des différentes typologies de connaissances, distinction entre les différentes sources de connaissances, etc.). De plus, ces extensions sont conçues de manière ad hoc, uniquement à travers des notations graphiques, sans définition formelle de la syntaxe abstraite. Certaines modifient même la sémantique de BPMN de manière non conforme (par exemple, l'intégration non autorisée d'éléments non-flux dans des séquences de flux), ce qui compromet leur validité au regard du standard BPMN.

En somme, l'analyse de la littérature révèle que la majorité des extensions de BPMN sont développées de manière peu structurée, souvent de façon « ad hoc », en se focalisant sur la syntaxe concrète sans démarche méthodologique claire. Cette approche limite leur compréhensibilité, leur réutilisation, ainsi que leur intégration dans les outils de modélisation. L'absence d'une méthode unifiée et rigoureuse pour guider la conceptualisation et la formalisation des extensions constitue un frein majeur à leur validité et à leur adoption. Ces constats motivent la proposition d'une méthodologie structurée, présentée dans la suite de ce manuscrit.

3. Objectifs de notre solution

Les travaux existants montrent une faible formalisation des premières étapes du développement des extensions BPMN, notamment l'analyse formelle des exigences spécifiques à un domaine et la conceptualisation. Ces étapes, pourtant fondamentales, manquent souvent d'une approche systématique et partagée, ce qui nuit à la clarté des modèles produits et limite la communication avec les experts métier. De plus, la phase de vérification d'équivalence entre les concepts du domaine et ceux du méta-modèle BPMN (cf. Braun and Schlieter, 2014 ; Braun et *al.*, 2015 ; 2016) est rarement détaillée, et repose sur des correspondances sémantiques peu

rigoureuses, compromettant la validité des extensions. Afin de répondre aux limites identifiées dans la littérature, nous proposons une méthodologie générique, cohérente et systématique, baptisée *EM-BPMN+X* (*Methodology for the Development of BPMN Plus Extensions*). Cette approche, fondée sur le paradigme du Design Science Research (DSR) (Hevner and Chatterjee, 2010), vise à guider l'ensemble du processus de développement d'extensions BPMN valides, (adaptées à des domaines spécifiques), de la spécification des exigences jusqu'à leur implémentation dans un outil de modélisation. EM-BPMN+X étend les travaux de Stroppi *et al*. (2011) et de Braun *et al*. (2014–2016), en mettant un accent particulier sur la phase de préparation (analyse des exigences, conception d'un modèle CDME) et la définition rigoureuse du méta-modèle de l'extension (BPMN+X), tant du point de vue de la syntaxe abstraite que concrète. Notre proposition introduit également une analyse sémantique approfondie entre les concepts du domaine cible et les éléments standards de BPMN, en s'appuyant notamment sur des ontologies de domaine et des patrons conceptuels ontologiques. L'objectif principal est de renforcer les capacités de BPMN 2.0.2 pour modéliser des processus métier complexes, flexibles, collaboratifs et à forte intensité de connaissances. EM-BPMN+X se veut une réponse aux besoins émergents de secteurs tels que la santé, la finance, les services publics, la cybersécurité, la logistique ou encore l'éducation. En offrant un cadre méthodologique structuré et extensible, cette approche améliore l'expressivité de BPMN, tout en facilitant la collaboration, la prise de décision fondée sur les connaissances, et la gestion efficace des risques dans des environnements organisationnels complexes.

4. Présentation de la méthode EM-BPMN+X

Le développement de EM-BPMN+X s'inscrit dans un projet de recherche à long terme, basé sur une approche centrée sur les exigences et intégrant les principes du DSR ((Hevner and Chatterjee, 2010) ; (Peffers *et al.*, 2018)). Plusieurs itérations et évaluations pratiques ont permis de raffiner les artefacts produits ((Ben Hassen *et al.*, 2017a ; 2017b ; 2022) ; (Ben Hassen et Gargouri, 2024)). Pour valider la méthodologie, nous l'avons appliquée au domaine des processus métier sensibles (SBP), ce qui a conduit au développement de l'extension *BPMN4SBP*. Cette dernière permet la spécification graphique des SBP dans une perspective d'identification et la gestion des connaissances cruciales. L'extension illustre la pertinence de EM-BPMN+X pour répondre aux exigences des processus complexes, hautement dynamiques, collaboratifs et à forte intensité de connaissances. La Figure 2 illustre la structure de la méthode EM-BPMN+X et son application pour la modélisation des SBP. Pour démontrer la validité de notre méthode et en faciliter la compréhension, chaque étape est illustrée par son application à la modélisation des SBP. En raison des contraintes de longueur de l'article, l'ensemble des ressources complémentaires – figures, tableaux, spécification complète du métamodèle et exemple de modélisation – est mis à disposition en ligne, structuré selon les étapes méthodologiques : [https://zenodo.org/records/15353258].

4.1. Phase 1 : Analyse du domaine

4.1.1 Étape 1.1: Analyse des exigences du domaine de modélisation

L'objectif de cette étape consiste à identifier les exigences (fonctionnelles) spécifiques au domaine de modélisation, en se basant sur une description des cas/scénarios d'utilisation ou sur une étude exhaustive de la littérature dans le domaine. Ces exigences permettent de spécifier les capacités et les caractéristiques de l'artefact visé du point de vue objectif. Par ailleurs, elles sont utiles pour prouver et valider la pertinence et l'adéquation du langage de modélisation BPMN 2.0.2 (OMG, 2013) pour satisfaire l'objectif visé et les différentes exigences identifiées. Subséquemment, ces exigences de domaine se traduisent en des concepts clés de domaine qui peuvent être spécifiés sous forme d'une ontologie noyau (de domaine).

Pour la modélisation des SBP, nous avons défini 14 exigences spécifiques à ces différentes dimensions (E1–E14)[2] (cf. Ben Hassen *et al.*, 2020 ; 2024b). Ces exigences servent de base pour la représentation des aspects statiques et dynamique de SBP.

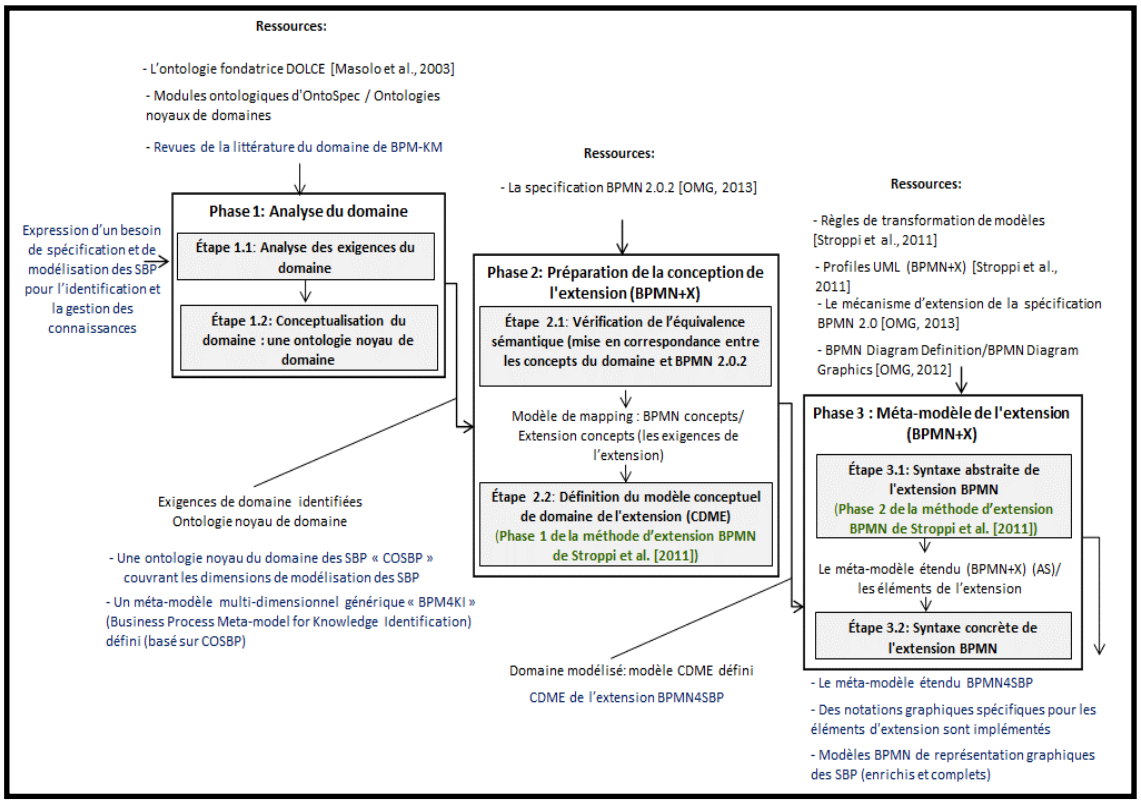


Figure 2. Méthode d'extension de BPMN 2.0.2 et son application pour la spécification des SBP

4.1.2. Étape 1.2: Conceptualisation du domaine : une ontologie noyau de domaine

Cette étape consiste à identifier les concepts, les propriétés, les relations, les règles et les contraintes du domaine. Ainsi, la réutilisation et l'intégration des ontologies s'avèrent utiles afin d'enrichir sémantiquement les différents modèles développés. Dans ce contexte, la conception d'une ontologie noyau de domaine (basée sur des ontologies fondatrices (e.g., DOLCE)) est la solution la mieux appropriée pour l'explicitation et la conceptualisation des connaissances du domaine de modélisation (et des extensions). En effet, l'ontologie spécifie un vocabulaire commun et précis du domaine qu'elle organise et formalise, afin de le rendre interprétable tant par les humains que par les machines. Ainsi, elle offre une sémantique formelle aux connaissances qu'elle explicite de sorte à permettre une application aisée du raisonnement. De ce fait, l'utilisation d'une ontologie est essentiellement liée pour répondre à deux objectifs essentiels : (i) représenter les connaissances d'un domaine ; et (ii) permettre de raisonner sur ces connaissances. Par ailleurs, la conception de l'ontologie noyau de domaine est très importante pour la deuxième phase de notre méthode d'extension de BPMN. En tant que référentiels offrant des définitions rigoureuses pour les concepts de domaine, cette ontologie permet de préparer le modèle conceptuel de domaine de l'extension « CDME » (Conceptual Domain Model of the Extension) en définissant une base vigoureuse pour l'étape de vérification de

[2]https://zenodo.org/records/15353258/files/Annexes%201&2_%20Classification%20des%20extensions%20de%20BPMN.pdf?download=1

l'équivalence sémantique des éléments. Les concepts clés et génériques de domaine définis rigoureusement doivent être examinés en ce qui concerne leur correspondance sémantique avec les éléments standards de BPMN.

Pour la conceptualisation du domaine des SBP, nous avons proposé une spécification conceptuelle rigoureuse intitulée COSBP (*Core Ontology of Sensitive Business Processes*) (Ben Hassen *et al.*, 2024a ; 2024b). Cette ontologie noyau de domaine a été élaborée selon une approche formelle multi-niveaux, reposant sur l'ontologie fondatrice DOLCE (Masolo *et al.*, 2003). COSBP réutilise et spécialise des concepts provenant de différentes ontologies noyaux de domaine lié aux BP, en vue de formaliser des concepts génériques pertinents pour les SBP. Elle offre une modélisation explicite, cohérente et rigoureuse des six dimensions fondamentales des SBP, à savoir les dimensions fonctionnelle, organisationnelle, comportementale, informationnelle, intentionnelle et connaissance, chacune représentée par une classe distincte de modules ontologiques[3].

4.2. Phase 2 : Préparation de la conception de l'extension (BPMN+X)

4.2.1. Étape 2.1: Vérification de l'équivalence sémantique

Cette étape consiste à comparer l'ensemble des concepts ontologiques du domaine cible aux éléments standards de BPMN, sur la base de définitions sémantiques. L'objectif est de déterminer les besoins réels en matière d'extension, en identifiant les éléments BPMN à réutiliser, à adapter ou à étendre. Cette vérification est essentielle pour garantir une utilisation appropriée de BPMN et éviter la création d'extensions superflues. Nous recommandons que tout concept d'extension possède une correspondance sémantique raisonnable avec un concept standard de BPMN. Lorsqu'une telle correspondance existe, il est préférable de spécifier l'élément à l'aide de l'infrastructure existante de BPMN, plutôt que d'introduire un nouveau concept par extension du *BaseElement*. Cela permet de limiter le nombre d'éléments étendus et de préserver la cohérence du langage. Pour guider cette analyse, nous adoptons les règles de correspondance définies par Braun *et al*. (2015) :

- **Équivalence** : le concept du domaine a une correspondance sémantique directe dans le méta-modèle BPMN, que ce soit un élément unique ou une combinaison valide d'éléments. Dans ce cas, aucune extension n'est requise et le concept est représenté comme un concept BPMN dans le modèle CDME (→ *CDME : BPMN Concept*).
- **Équivalence conditionnelle** : il n'existe pas de correspondance explicite, mais une réflexion contextuelle permet d'évaluer si un élément BPMN peut représenter le concept du domaine. Il convient alors de justifier cette mise en correspondance, ou d'expliquer pourquoi elle n'est pas envisageable. Cette analyse est d'autant plus importante que le méta-modèle BPMN propose des éléments à la sémantique parfois très large (OMG, 2013). Partant de ce fait, le concept est traité soit comme un concept équivalent, soit comme étant un concept non équivalent.
- **Absence d'équivalence** : aucune correspondance ne peut être établie, ce qui peut s'expliquer par : (1) l'absence totale du concept dans BPMN → *CDME : Extension Concept* ; (2) l'absence d'une relation entre deux concepts → *CDME : Association entre concepts* ; (3) l'absence d'attributs spécifiques à un concept → *CDME : Propriété spécifique d'un concept*.

En complément de ces règles, nous proposons deux types d'équivalences sémantiques enrichies (Ben Hassen *et al.*, 2017a ; 2022) :

- **Équivalence par composition** : un concept du domaine peut être représenté par une composition cohérente d'éléments standards de BPMN.

[3] https://zenodo.org/records/15353258/files/EM-BPMN+X_%C3%89tape%201-2_%20Conceptualisation%20du%20domaine%20-%20une%20ontologie%20noyau%20de%20domaine%20(COSBP).pdf?download=1

- **Équivalence par spécification** : le concept du domaine est une spécialisation d'un élément BPMN existant, par l'ajout de sémantique ou de propriétés propres au domaine étudié.

Les résultats de cette vérification peuvent être synthétisés dans un tableau d'analyse ou un modèle de mappage spécifique, servant de base à la préparation du modèle CDME et à la définition des stéréotypes correspondants (Ben Hassen *et al.*, 2022). Un extrait illustrant la vérification d'équivalence entre l'ontologie COSBP et les concepts de BPMN 2.0.2, ainsi que la dérivation des concepts du modèle CDME de l'extension BPMN4SBP, est également disponible en ligne[4].

4.2.2. Étape 2.2: Définition du modèle conceptuel de domaine de l'extension (Conceptual Domain Model of the Extension- CDME)

Cette étape correspond à la modélisation du domaine de l'extension (BPMN+X), également désignée comme la construction du modèle CDME, tel que défini dans la phase 1 de la méthode d'extension BPMN proposée par Stroppi *et al.* (2011). Le CDME (Conceptual Domain Model of the Extension) a pour objectif de représenter l'ensemble des concepts spécifiques au domaine cible devant être intégrés à BPMN, ainsi que leurs relations sémantiques avec les éléments du méta-modèle BPMN existant. Ce modèle est élaboré sous forme d'un diagramme de classes UML, indépendamment des contraintes imposées par le mécanisme formel d'extension de BPMN. Cette indépendance permet de se concentrer avant tout sur la sémantique des concepts, sans se restreindre aux considérations techniques (imposées par le mécanisme d'extension de BPMN) liées à leur future implémentation dans BPMN. Dans ce modèle, les classes sont typées en amont selon deux catégories : (i) « *BPMN Concepts* » (concepts standards du langage BPMN pouvant être réutilisés) ; (ii) « *Extension Concepts* » (nouveaux concepts propres au domaine cible, absents du méta-modèle BPMN, et qui nécessitent une extension). Le modèle CDME constitue ainsi une base conceptuelle solide pour la définition rigoureuse de la syntaxe abstraite de l'extension BPMN+X. Il permet d'assurer une cohérence entre les exigences du domaine et les capacités d'expression offertes par le langage étendu (Ben Hassen et al., 2022 ; Ben Hassen and Gargouri, 2024). Un extrait illustrant le modèle CDME de l'extension BPMN4SBP est disponible en ligne[5].

4.3. Phase 3 : Méta-modèle de l'extension (BPMN+X)

4.3.1. Étape 3.1: Syntaxe abstraite de l'extension BPMN : le méta-modèle BPMN+X (AS)

Cette étape vise à concevoir le méta-modèle étendu (BPMN+X) à partir du modèle CDME préalablement élaboré. La sémantique et la syntaxe abstraite des éléments de BPMN+X s'appuient sur la spécification du mécanisme d'extension de BPMN. Ce méta-modèle est défini sous forme d'un profil UML, composé de plusieurs stéréotypes tels que *ExtensionModel*, *BPMNElement*, *ExtensionElement*, *ExtensionDefinition*, *ExtensionEnum*, *BPMNEnum* et *ExtensionRelationship* (Stroppi *et al.*, 2011). Ce profil BPMN+X permet ainsi de représenter graphiquement les extensions apportées à BPMN.

Le modèle d'extension (*ExtensionModel*) constitue le conteneur principal regroupant tous les éléments définissant une extension BPMN. Il repose sur plusieurs classes stéréotypées, chacune jouant un rôle spécifique dans la représentation des éléments du méta-modèle BPMN 2.0.2 et des concepts ajoutés. Les classes `BPMNElement` représentent les éléments standards du méta-modèle BPMN. Les classes `BPMNEnum` et `ExtensionEnum` permettent de définir

[4] https://zenodo.org/records/15353258/files/EM-BPMN+X_Etape%202-1-V%C3%A9rification%20d'%C3%A9quivalence%20(entre%20les%20concepts%20de%20COSBP%20et%20de%20BPMN%202.0.2).pdf?download=1

[5] https://zenodo.org/records/15353258/files/EM-BPMN+X_2tape%202-2_Le%20mod%C3%A8le%20CDME%20de%20l%E2%80%99extension%20BPMN4SBP%20(un%20extrait).pdf?download=1

des ensembles de littéraux, respectivement issus du méta-modèle BPMN 2.0.2 et du modèle d'extension. La classe `ExtensionElement` est utilisée pour introduire un nouveau type d'élément absent du méta-modèle BPMN. Quant à `ExtensionDefinition`, elle regroupe un ensemble nommé d'attributs ajoutés à un élément BPMN existant, conformément au concept du même nom dans la spécification BPMN (OMG, 2013). Les attributs d'extension, correspondant à l'élément `ExtensionAttributeDefinition` du méta-modèle BPMN, sont modélisés via la métaclasse `Property` d'UML. Ils sont définis comme des propriétés UML appartenant à une `ExtensionDefinition` ou accessibles via des associations. Ces propriétés peuvent être typées comme suit : `BPMNElement`, `ExtensionElement`, `BPMNEnum`, `ExtensionEnum`, ou un type primitif UML. Un élément particulier, `ExtensionRelationship`, est utilisé pour établir un lien conceptuel entre un élément BPMN existant (`BPMNElement`) et une `ExtensionDefinition` censée l'étendre. Ce mécanisme, bien qu'inopérant sur la structure de l'extension générée, a une vocation conceptuelle. Il sert à guider la définition et la compréhension des extensions, notamment lors de l'adaptation du méta-modèle BPMN à un nouveau domaine. Par ailleurs, Stroppi *et al*. (2011) proposent un ensemble de contraintes OCL à intégrer au profil BPMN+X pour garantir la cohérence des extensions avec les mécanismes d'extension définis par la spécification BPMN.

Démarche de transformation du CDME en BPMN+X. Pour générer un modèle BPMN+X à partir du modèle conceptuel de domaine (CDME), nous nous appuyons sur la démarche en deux étapes proposées par Stroppi *et al*. (2011), visant à produire un profil BPMN+X valide :

1. **Création et peuplement du modèle BPMN+X** : Cette étape consiste à instancier le modèle `ExtensionModel` en intégrant : (i) Les `BPMNElement` et `BPMNEnum` correspondant aux *BPMN Concepts* identifiés dans le CDME ; (ii) Les `ExtensionElement`, `ExtensionEnum`, et `ExtensionDefinition` correspondant aux *Extension Concepts*.
2. **Application des règles de transformation** : Quinze règles définies par Stroppi *et al.* (2011) sont appliquées pour traduire les concepts d'extension du CDME en éléments BPMN+X. Ces règles se basent sur : (i) l'analyse des propriétés de classes et des associations (R1a, R1b, R2a, R2b, R2c, R3, R4a, R4b, R4c ; (ii) les relations de généralisation entre concepts (R5, R6, R7a, R7b, R8a, R8b) (cf. Ben Hassen *et al.*, 2022 ; Ben Hassen and Gargouri, 2024). Elles garantissent une transformation rigoureuse et conforme aux mécanismes d'extension de BPMN 2.0.2 (OMG, 2013).

Cette démarche consolidée est à la fois robuste et générique, car elle permet de générer un modèle BPMN+X valide, même à partir de CDME hétérogènes. Elle garantit ainsi une formalisation cohérente et conforme des extensions du langage BPMN. Deux exemples illustrant l'application de ces règles à partir des *Extension Concepts* du CDME sont présentés en ligne[6]. En tant que profil UML, le méta-modèle BPMN4SBP se compose de plusieurs stéréotypes : `BPMN Element`, `Extension Element`, `Extension Definition`, `Extension Relationship` et `Extension Enum`.[7]

4.3.2. Étape 3.2: Syntaxe concrète de l'extension BPMN+X

En complément de la syntaxe abstraite, il est nécessaire de définir une syntaxe concrète normalisée, conformément aux directives de la spécification BPMN. Cette syntaxe graphique facilite l'échange des modèles et leur intégration dans des outils BPMN existants. Il s'agit ici de

[6] https://zenodo.org/records/15353258/files/EM-BPMN+X_%C3%89tape%203-1_Application%20des%20r%C3%A8gles%20de%20transformation%20(CDME--BPMN4SBP).pdf?download=1

[7] https://zenodo.org/records/15353258/files/EM-BPMN+X_Etape%203-1_D%C3%A9finition%20de%20la%20syntaxe%20abstraite%20de%20l%E2%80%99extension%20BPMN4SBP%20(un%20extrait).pdf?download=1

spécifier de manière systématique la représentation graphique des nouveaux éléments introduits par BPMN+X. Pour cela, nous nous appuyons sur la spécification Diagram Definition (DD) de l'OMG (2012), en particulier sur le package Diagram Graphics (DG), instanciable via le sous-ensemble BPMN-DG. Ce dernier permet une représentation graphique conforme et extensible des éléments ajoutés. La réussite de cette démarche repose aussi sur la mise en œuvre de l'extension dans un outil BPMN adapté. Plusieurs plateformes peuvent être envisagées (Activiti, Bizagi, Cubetto, BPMN2 Modeler), à condition de répondre aux critères suivants : (i) compatibilité complète avec BPMN 2.0; (ii) développement en Java pour faciliter l'adaptation ; (iii) nature open source garantissant accessibilité et évolutivité; (iv) extensibilité graphique permettant l'ajout intuitif de nouveaux éléments ; (v) documentation claire, accessible aux développeurs non experts; (vi) accessibilité aux utilisateurs finaux sans compétences techniques avancées. Ainsi, cette étape garantit à la fois la cohérence formelle et l'opérabilité de l'extension dans un environnement orienté utilisateur. Pour la représentation graphique des SBP, Ben Hassen *et al.* (2019) ont développé BPMN4SBP-Modeler, un plug-in Eclipse basé sur BPMN2 Modeler. Cette extension permet de modéliser explicitement les dimensions fonctionnelle, organisationnelle, informationnelle, comportementale, intentionnelle et de connaissance, souvent négligées dans les outils classiques. Il vise à soutenir une modélisation multidimensionnelle des SBP tout en restant compatible avec la notation BPMN enrichie proposée. Un extrait de l'extension développée est disponible en ligne[8].

5. Démonstration et évaluation

L'extension BPMN4SBP répond aux exigences de modélisation des processus sensibles à forte intensité de connaissances (SBP) dans les environnements de KM. La méthode EM-BPMN+X garantit la cohérence entre l'analyse du domaine, la conceptualisation ontologique et l'implémentation du méta-modèle validé. Cette approche a été appliquée et évaluée dans le domaine médical, notamment à l'Association de Sauvegarde des Handicapés Moteurs de Sfax (ASHMS)**,** pour modéliser la prise en charge précoce des enfants atteints d'une infirmité motrice cérébrale (IMC) (Ben Hassen *et al.*, 2017a ; 2017b ; 2022 ; 2024a ; Ben Hassen & Gargouri, 2024). Un extrait du modèle relatif au « Processus d'évaluation initiale (neuro-moteur/neuro-musculaire et neuro-développemental) d'un enfant IMC » est présenté en ligne, illustrant l'utilisation de l'outil BPMN4SBP Modeler. L'évaluation a impliqué des experts BPM-KM, des ingénieurs en systèmes d'information et des professionnels de santé de l'ASHMS. Elle a été réalisée à travers des entretiens structurés et des sessions pratiques, avec les critères suivants : (i) *Compréhension et complétude* – Capacité à modéliser clairement les dimensions des SBP et à identifier les connaissances cruciales dans les processus. (ii) *Utilité de BPMN4SBP-Modeler* **–** Efficacité dans l'analyse et la représentation des SBP, en particulier dans des domaines sensibles comme la santé. (iii) *Facilité d'utilisation* **–** Intuitivité, clarté fonctionnelle et accessibilité pour les utilisateurs non experts. (iv) *Concordance méthode-outil* – Cohérence entre les concepts du méta-modèle et les fonctionnalités de l'éditeur BPMN4SBP. Les résultats quantitatifs (Figure 3) montrent des scores élevés sur tous les critères, confirmant une perception très positive parmi les participants. BPMN4SBP-Modeler s'avère être un outil efficace et bien aligné, particulièrement adapté pour la modélisation des SBP dans des environnements KM, comme celui de la santé. Les scores de concordance indiquent que l'outil traduit avec succès les principes théoriques de modélisation des SBP en une solution fonctionnelle et accessible.

En conclusion, les résultats confirment que BPMN4SBP améliore significativement l'expressivité de BPMN, tout en facilitant la modélisation de processus collaboratifs,

[8]https://zenodo.org/records/15353258/files/Syntaxe%20concr%C3%A8te%20de%20l'extension%20BPMN4SBP%20et%20son%20application%20pour%20la%20mod%C3%A9lisation%20d'un%20SBP.pdf?download=1

dynamiques, complexes et riches en connaissances. La validation de cette méthode atteste de son efficacité pour la gestion des SBP, en constituant un outil décisionnel précieux et un levier stratégique pour la gestion des processus métier contemporains.

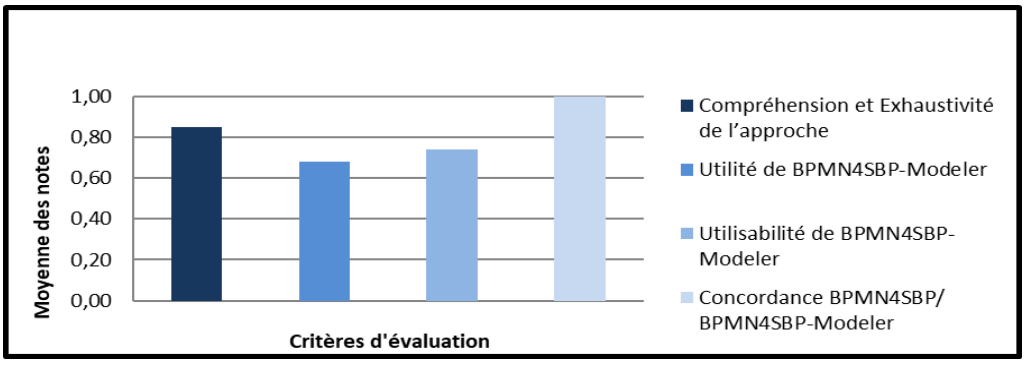


Figure 3. Résultat de l'évaluation des contributions

6. Conclusion

Cette recherche aborde les problèmes des « Trois Fit » – *Vertical*, *Horizontal* et *Transversal* – qui limitent l'intégrité, la flexibilité et l'interopérabilité des Systèmes d'Information d'Entreprise (EIS). Pour y répondre, nous proposons la méthode EM-BPMN+X, fondée sur la *Design Science Research* (Peffers et al., 2018), permettant de concevoir des extensions valides de BPMN 2.0.2 adaptées à des domaines spécifiques. Cette méthode comble l'écart entre l'analyse métier et la conception de méta-modèles BPMN, en s'appuyant sur deux phases principales : (1) L'intégration d'une ontologie de domaine, garantissant une représentation claire et partagée des concepts métier ; (2) Une comparaison sémantique avec les éléments standards de BPMN, pour n'introduire que les extensions réellement nécessaires. La notation BPMN+X enrichit ainsi BPMN 2.0.2 pour mieux représenter des processus complexes, flexibles, dynamiques, collaboratifs et à forte intensité de connaissances, caractéristiques des environnements modernes. Au-delà de ses apports théoriques, l'extension BPMN+X offre des bénéfices concrets dans divers domaines : elle facilite l'automatisation des BP, améliore l'utilisation des ressources et réduit les coûts opérationnels. Elle se révèle particulièrement pertinente dans des secteurs comme la santé. Elle met à disposition des outils pratiques pour modéliser et piloter des processus riches en connaissances, renforçant ainsi la prise de décision, la collaboration et l'efficacité organisationnelle. Alignée sur les standards BPMN, cette extension s'intègre aisément aux systèmes de gestion des processus (BPMS), ce qui simplifie son déploiement en contexte réel. Sur le plan éducatif, elle constitue un support d'apprentissage efficace pour l'enseignement de la gestion des BP, de KM et des systèmes d'information, notamment à travers des cas concrets telles que celles issues du secteur de la santé. Sur le plan sociétal, BPMN+X contribue à améliorer l'identification et le partage des connaissances, la localisation de l'information et la qualité des décisions, tout en favorisant l'apprentissage organisationnel et la prise de décision collaborative. Cette recherche ouvre également de nombreuses perspectives. Des travaux futurs pourront enrichir la méthode EM-BPMN+X en intégrant des ontologies de domaine en OWL, facilitant ainsi leur usage dans les EIS et les applications du Web sémantique. L'adoption de l'approche Model-Driven Engineering (MDE) permettra également d'automatiser la transformation des modèles BPMN+X en spécifications exécutables. La généralisation et la validation de la méthode dans d'autres domaines, tels que la finance, la logistique, l'éducation, renforceront sa portée. Enfin, des recherches pourront explorer son application dans des contextes critiques comme la gestion de crise ou la définition de politiques publiques, où réactivité et adaptabilité sont cruciales. À terme, l'objectif est de renforcer la flexibilité, l'agilité et les capacités d'automatisation de BPMN4SBP, pour en garantir une adoption élargie et pertinente dans la gestion des BP.

Bibliographie

Abouzid, I. and Saidi, R. (2019). Proposal of BPMN extensions for modelling manufacturing processes. In the *5th International Conference on Optimization and Applications*, pp. 1-6.

Ben Hassen, M., Turki M. and Gargouri, F. (2017a). Extending sensitive business process modeling with functional dimension for knowledge identification". In *Proceedings of the 14th International Conference on e-Business (ICE-B 2017*), Madrid, Spain, Vol. 2, pp. 38-51.

Ben Hassen, M., Turki M. and Gargouri, F. (2017b). BPMN4KM: Design and Implementation of a BPMN Extension for Modeling the Knowledge Perspective of Sensitive Business Processes. *Journal of Procedia Computer Science*, Vol. 121, pp.1119-1134.

Ben Hassen, M., Turki M. and Gargouri, F. (2022). Extending BPMN 2.0 Models with Sensitive Business Process Aspects". In the *26th International Conference on Knowledge Based and Intelligent Information and Engineering Systems (KES'2022)*, Italy. Vol 207, 2022, pp. 2968-2979

Ben Hassen, M and Gargouri, F. (2024). Graphical Specification of Sensitive Business Processes. *Procedia Computer Science*, Vol. 237, p.p. 96-106.

Ben Hassen, M., Turki M. and Gargouri, F. (2019). A Multicriteria Evaluation Approach for Selecting a Sensitive Business Process Modeling Language for Knowledge Management. *Journal on Data Semantics (JODS)*, Vol.8 No.3, pp. 157-202.

Ben Hassen, M., Zahaf. S., and Gargouri, F. (2024a). Resolving the "three-fit" problems in enterprise information systems: a design science approach with ontological solutions. *Business Process Management Journal*.

Ben Hassen, M., Turki, M., and Gargouri, F. (2024b). Conceptual analysis of sensitive business processes. *Business Process Management Journal*, 30(5), 1501-1540.

Ben Said, I., Chaâbane, M. A., Andonoff, E. and Bouaziz, R. (2018), "BPMN4VC-modeller: easy-handling of versions of collaborative processes using adaptation patterns". International Journal of Information Systems and Change Management, 10(2), pp. 140-189.

Betke and Seifert (2017), "BPMN for disaster response processes – a methodical extension", INFORMATIK Conference, pp. 1311-1324.

Braun, R., Schlieter, H., Burwitz, M. and Esswein, W. (2015). Extending a Business Process Modeling Language for Domain-Specific Adaptation in Healthcare. in: Thomas. O.; (Hrsg.): *Internationalen Tagung Wirtschaft (WI 2015)*, Osnabrück, S, pp. 468-481.

Braun, R., Schlieter, H., Burwitz, M. and Esswein, W. (2016). BPMN4CP revised – extending BPMN for multi-perspective modeling of clinical pathways. *49th Hawaii International Conference on System Sciences (HICSS)*, Koloa, HI, USA, pp. 3249-3258.

Cartelli, V., Di Modica, G. and Tomarchio, O. (2016), "Extending the BPMN specification to support cost-centric simulations of business processes". IC3K, pp. 492-514.

Carvalho, L.P., Cappelli, C. and Santoro, F.M. (2018), "AO-BPM 2.0: aspect-oriented business process modeling", Business Process Management Workshops, pp. 719-731.

Chergui, M. E. A. and Benslimane, S. M. (2020), "Towards a BPMN Security Extension for the Visualization of Cyber Security Requirements". *IJ of Technology Diffusion (IJTD)*, 11(2), pp. 1-17.

Dukaric, R. and Juric, M.B. (2018), BPMN extensions for automating cloud environments using a two-layer orchestration approach, J. of Visual Languages and Computing, 47, pp. 31-43.

Fournier-Morel, X., Grojean, P., Rognon, C. (2008), SOA le Guide de l'Architecture du SI, Dunod, Paris.

Hevner, A. and Chatterjee, S. (2010), "Design Research in Information Systems": Theory and Practice. Springer, New York, 320.

Heguy, X., Zacharewicz, G., Ducq, Y.. and Vallespir, B. (2019). A performance measurement extension for BPMN: one step further quantifying interoperability in process model. Enterprise Interoperability VIII, Proceedings of the I-ESA Conferences, pp. 333-345

Intrigila, B., Della Penna, G., & D'Ambrogio, A. (2021). A lightweight BPMN extension for business process-oriented requirements engineering. Computers, 10(12), 171.

Louar, F., Zarour, K. and Benmerzoug, D. (2018). Modelling Business Processes for Outsourcing into the Fog and Cloud Computing. In SIMPDA, pp. 18-31.

Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A., Schneider L. and Horrocks, I. (2003). The WonderWeb Library of Foundational Ontologies and the DOLCE ontology. WonderWeb Deliverable D18, Final Report (version 1.0, 31-12-2003).

Neumann, J., Franke, S., Rockstroh, M., Kasparick, M. and Neumuth, T. (2019). Extending BPMN 2.0 for intraoperative workflow modeling. *International Journal of Computer Assisted Radiology and Surgery*, Vol. 14 No. 8, pp. 1-11.

Pufahl, L., Zerbato, F., Weber, B., & Weber, I. (2022). BPMN in healthcare: Challenges and best practices. Information Systems, 107, 102013.

OMG (2012), *Diagram Definition (DD)*, Version 1.0. http://www.omg.org/spec/DD/1.0/PDF/.

OMG (2013), *Business process model and notation (BPMN)*, available at: http://www.omg.org/spec/ BPMN/2.0.2 (accessed 10 April 2021).

OMG (2014). *Meta Object Facility (MOF)*, Version 2.4.2. http://www.omg.org/spec/MOF/.

Onggo, B., Proudlove, N., D'Ambrogio, S., Calabrese, A., Bisogno, S. and Levialdi Ghiron, N. (2018). A BPMN extension to support discrete-event simulation for healthcare applications: an explicit representation of queues, attributes and data-driven decision points. *Journal of the Operational Research Society*, Vol. 69 No. 5, pp. 788-802.

Peffers, K., Tuunanen, and Niehaves, B. (2018), "Design Science Research Genres: Introduction to the Special Issue on Exemplars and Criteria and Applicable Design Science Research". European Journal of Information Systems, Vol.27(2), 129.

Polančič, G. (2020), "BPMN-L: A BPMN extension for modeling of process landscapes". Computers in Industry, 121, 103276

Polderdijk, M., Vanderfeesten, I., Erasmus, J., Traganos, K., Bosch, T., Rhijn, G. and Fahland, D. (2018). A visualization of human physical risks in manufacturing processes using BPMN. *Business Process Management Workshops*, pp. 732-744.

Ramos-Merino, M., Santos-Gago, J. M., Álvarez-Sabucedo, L. M., Alonso-Roris, V. M. and Sanz-Valero, J. (2019), "BPMN-E 2: a BPMN extension for an enhanced workflow description". Software & Systems Modeling, 18(4), pp. 2399-2419

Santra, D. and Choudhury, S. (2018). C-BPMN: A Context Aware BPMN for Modeling Complex Business Process. arXiv preprint arXiv:1806.01333.

Skouti, T., Seiger, R., Furrer, F. J., & Strahringer, S. (2024). RBPMN: the value of roles for business process modeling. Software and Systems Modeling, 1-32.

Stroppi, L.J.R., Chiotti, O. And Villarreal, P.D. (2011), "Extending BPMN 2.0: Method and tool support". In Business Process Model and Notation, Dijkman, R., Hofstetter, J., Koehler, J. (eds.) BPMN 2011. LNBIP, vol. 95, pp. 59–73. Springer, Heidelberg.

Strutzenberger, D., Mangler, J. & Rinderle-Ma, S. (2024), "Evaluating BPMN Extensions for Continuous Processes Based on Use Cases and Expert Interviews". Bus Inf Syst Eng.

Szelągowski, M., Biernacki, P., Berniak-Woźny, J., & Lipinski, C. R. (2022), "Proposal of BPMN extension with a view to effective modeling of clinical pathways". Business Process Management Journal, 28(5/6), 1364-1390.

Vogel, J., Zobel, B., Jannaber, S. and Thomas, O. (2018). BPMN4SGA: a BPMN extension for smart glasses applications to enable process visualisations. In *Workshops der INFORMATIK 2018-Architekturen, Sicherheit und Nachhaltigkeit*. Verlag GmbH.

Zarour, K., Benmerzoug, D., Guermouche, N., & Drira, K. (2019). A BPMN extension for business process outsourcing to the cloud. In *New Knowledge in Information Systems and Technologies*: Vol 1. Springer International Publishing, pp. 833-843.

Détection du Mensonge : Revue de Littérature sur l'Analyse des Expressions Faciales et le Machine Learning

Monica Sen, Rébecca Deneckère

Centre de Recherche en Informatique
Université Paris 1 Panthéon-Sorbonne
rebecca.deneckere@univ-paris1.fr

REFERENCE DE L'ARTICLE INTERNATIONAL Cet article est une synthèse de l'article : Monica Sen, Rébecca Deneckère: Unmasking Lies: A Literature Review on Facial Expressions and Machine Learning for Deception Detection. KES 2024: 1925-1935

1. Introduction

Cet article explore l'utilisation des expressions faciales par machine learning pour détecter le mensonge, en mettant l'accent sur les défis liés à la qualité des données nécessaires au développement de modèles efficaces. Depuis longtemps, la question de savoir si un comportement observable ou des preuves permettent de distinguer un menteur d'une personne honnête fascine les chercheurs. Les méthodes traditionnelles, comme le polygraphe, présentent des limites importantes (Burzo et al, 2018) : intrusivité, failles dans la détection, et biais humains. De plus, ces outils ne parviennent souvent pas à distinguer les menteurs des honnêtes, et les résultats peuvent être erronés, exonérant des coupables ou accusant des innocents. L'avènement des technologies d'intelligence artificielle, offre une alternative prometteuse.

Le comportement humain, et en particulier le mensonge, est perçu comme un phénomène complexe influencé par la cognition, les émotions et les actions (Bhatt et al, 2023). L'intelligence artificielle, en se basant sur de vastes ensembles de données, a permis de mieux comprendre et modéliser les comportements cognitifs et émotionnels humains. Bien que l'apprentissage automatique puisse analyser des expressions faciales, telles que les micro-expressions et les macro-expressions, pour détecter des indices de mensonge, ces techniques sont encore largement limitées par la diversité des comportements humains et les défis posés par les biais dans les ensembles de données.

2. Analyse

Ce travail essaie de répondre à la question de recherche suivante : « **Le machine learning peut-il aider à détecter le mensonge à partir des expressions faciales ?** ». Nous avons sélectionné huit expérimentations particulièrement pertinentes sur le sujet pour pouvoir les analyser et les comparer, à la fois sur les données utilisées, l'extraction des caractéristiques de mensonges, la classification des résultats et l'évaluation des modèles proposés.

Les expressions faciales sont classifiées en deux types : les macro-expressions, et les micro-expressions. Les recherches récentes utilisent ces systèmes en combinaison avec des modèles d'apprentissage automatique pour identifier les indices faciaux associés au mensonge.

Les modèles existants présentent des résultats contradictoires. Bien que certains montrent une précision prometteuse, cette précision est souvent limitée à des contextes spécifiques, avec un nombre restreint de participants dans les études. Cela soulève la question de leur généralisation à des situations réelles, où les comportements de mensonge peuvent varier en fonction du contexte culturel, de l'expérience personnelle et de l'état émotionnel des individus. De plus, les modèles basés uniquement sur les expressions faciales sont confrontés à des limites importantes, car de nombreux mensonges ne sont pas nécessairement reflétés par des expressions faciales. D'autres indices, comme le langage corporel ou des incohérences verbales, peuvent également jouer un rôle crucial dans la détection.

3. Conclusion

L'une des principales recommandations pour améliorer la robustesse et la fiabilité des modèles est d'élargir la diversité des ensembles de données, afin d'inclure des contextes culturels et des situations de la vie réelle multiples. L'adaptation des modèles à de nouveaux contextes et l'amélioration de leur interprétabilité sont également des objectifs importants pour accroître la transparence et la confiance dans les systèmes de détection. Enfin, des préoccupations éthiques doivent être prises en compte, notamment pour éviter les biais et garantir le respect de la vie privée des individus. Des améliorations dans la collecte de données, l'interprétabilité des modèles et le respect des principes éthiques sont nécessaires pour rendre ces systèmes fiables et applicables dans des contextes variés.

Bibliographie

Bhatt, Priya, Amanrose Sethi, Vaibhav Tasgaonkar, Jugal Shroff, Isha Pendharkar, Aditya Desai, Pratyush Sinha et al. (2023) Machine learning for cognitive behavioral analysis: datasets, methods, paradigms, and research directions. Brain informatics 10(1): 18.

Burzo, Mihai, Mohamed Abouelenien, Veronica Perez-Rosas, and Rada Mihalcea (2018) Multimodal deception detection. The Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, and Detection of Emotion and Cognition-Volume 2, pp. 419-453

Une Revue Systématique de la Littérature sur les Techniques d'Affective Computing pour la Détection du Stress au Travail

Défis et Perspectives, de la Collecte des Données à la Détection du Stress

Iris Mezieres[1], Abir Gorrab[2], Rébecca Deneckère[1], Nourhène Ben Rabah[1] and Bénédicte Le Grand[1]

1. Centre de Recherche en Informatique
Université Paris 1 Panthéon-Sorbonne, Paris, France
{Rebecca.Deneckere,Nourhene.Ben-Rabah,Benedicte.Le-Grand}@univ-paris1.fr
2. RIADI Laboratory
National School of Computer Science, University of Manouba, Tunisia
Abir.Gorrab@riadi.rnu.tn

REFERENCE DE L'ARTICLE INTERNATIONAL Cet article est une synthèse de l'article :
Iris Mezieres, Abir Gorrab, Rébecca Deneckère, Nourhène Ben Rabah, Bénédicte Le Grand: A Systematic Literature Review on Affective Computing Techniques for Workplace Stress Detection - Challenges, Future Directions, from Data Collection to Stress Detection. ICCCI (CCIS Volume 1) 2024: 44-56

1. Introduction

Dans un monde où le travail occupe une place centrale dans la vie quotidienne, son impact sur le bien-être des individus est indéniable. Le stress au travail est devenu un sujet d'attention croissante en raison de ses conséquences profondes sur la santé des employés et la performance des entreprises. La détection et la gestion du stress au sein des environnements professionnels constituent ainsi un enjeu essentiel pour favoriser un cadre de travail sain. L'émergence de nouvelles technologies, notamment dans le domaine de l'affective computing, ouvre des perspectives prometteuses en matière d'évaluation du stress. L'affective computing repose sur l'utilisation de divers outils informatiques permettant d'analyser le comportement et les émotions humaines. Cet article présente une revue systématique de la littérature (SLR) portant sur les recherches existantes en matière d'évaluation du stress au travail à l'aide des technologies d'affective computing.

2. Analyse

Cette revue vise à répondre à la question suivante : "**Comment les technologies d'affective computing sont-elles utilisées pour améliorer la mesure du stress des employés au travail ?**" Nous avons suivi la méthode préconisée par (Kitchenham et Charters, 2007) et nous sommes concentrés sur un total de 35 articles pertinents. Chaque article a ensuite été étudié sous le prisme de quatre aspects clés :

- les domaines d'application : ceux-ci peuvent être statiques ou dynamiques,

- la collecte des données : les données sont extrêmement diverses (physiologiques, physiques, de comportement, etc.) et peuvent être collectées de différentes manières (par des senseurs portés ou manipulés par les utilisateurs, par des senseurs inclus dans les bureaux, par des senseurs ou des caméras plus traditionnels ou par des enquêtes),

- l'analyse des données : cette analyse peut être faite par des solutions basées sur l'intelligence artificielle (machine learning, deep learning) ou par d'autres types de solutions (des modèles mathématiques par exemple), et

- les défis associés : plusieurs défis sont identifiables dans les articles de notre corpus, comme la collection éthique de données, l'anonymisation des données, le stockage sécurisé des données, l'accessibilité des résultats, l'efficacité de l'identification du stress ou encore l'implémentation en environnement réel.

3. Conclusion

L'utilisation de l'affective computing pour mesurer le stress au travail présente un grand potentiel, mais elle est confrontée à plusieurs obstacles majeurs. En particulier, l'interprétabilité des résultats reste un défi important, notamment lorsqu'on utilise des méthodes d'apprentissage profond appliquées à l'analyse des expressions faciales. Les auteurs envisagent, dans leurs futurs travaux, de mener des expériences en contexte réel en exploitant des algorithmes d'apprentissage profond pour détecter les émotions négatives, comme le stress et le désespoir, à partir des expressions faciales. Ces expérimentations seront menées sur de larges populations afin de proposer des solutions concrètes et efficaces aux travailleurs en situation de stress. En conclusion, cette revue systématique met en lumière l'intérêt grandissant pour l'utilisation des technologies d'affective computing dans l'évaluation du stress au travail. Cependant, plusieurs défis techniques et éthiques doivent encore être relevés pour permettre une application fiable et éthique de ces technologies en entreprise. L'avenir de ces recherches repose sur des tests en conditions réelles et sur l'amélioration des techniques d'analyse des émotions afin d'offrir des solutions adaptées aux besoins des employés et des employeurs.

Bibliographie

Kitchenham, B., Charters, S. (2007) Guidelines for performing systematic literature reviews in software engineering

Une approche hybride combinant Markov, HMM et RNN pour détecter les blocages dans l'apprentissage de la programmation

Grota Abdelkader [2], Mohammed Erritali [1,2], Patrick Etcheverry [1], Thierry Nodenot [1]

1.T2I, Laboratoire LIUPPA, IUT de Bayonne, France.
2.Laboratoire Data4Earth, Faculté des Sciences et Techniques, Béni Mellal Maroc

RÉSUMÉ. *Comprendre le processus d'apprentissage en programmation est un défi complexe en raison de la nature séquentielle et multidimensionnelle des interactions des étudiants avec les environnements numériques. Cette étude vise à analyser les journaux d'activité des étudiants pour détecter les difficultés rencontrées et proposer des interventions pédagogiques adaptées. Nous présentons une approche hybride combinant les Chaînes de Markov, les Modèles Cachés de Markov (HMM) et les Réseaux Neuronaux Récurrents (RNN) avec mécanisme d'Attention, afin d'exploiter les dimensions comportementale, cognitive et structurelle de l'apprentissage.La méthodologie consiste à extraire des caractéristiques séquentielles des journaux d'activité et à modéliser les transitions entre actions pour identifier des schémas tels que les cycles d'exploration, d'hésitation et de blocage. Ces schémas sont intégrés dans un cadre unifié pour prédire les états d'apprentissage des étudiants et détecter les moments critiques de difficulté. Le modèle hybride a été validé sur des données réelles issues de 5 étudiants ayant indiqué avoir réussi à terminer leurs travaux pratiques, avec un total de 26 enregistrements de traces. L'exercice visait à implémenter un tri à bulles sur des entiers, puis à appliquer cette méthode sur d'autres structures de données.Les résultats montrent que le modèle hybride proposé surpasse les approches traditionnelles, avec une précision de 93,5% et une réduction significative des faux positifs dans la détection des blocages. L'analyse multidimensionnelle permet de mieux comprendre les comportements et les trajectoires d'apprentissage des étudiants.Cette recherche ouvre la voie au développement de plateformes éducatives intelligentes capables de fournir un feedback personnalisé, favorisant ainsi une meilleure réussite et un engagement accru des étudiants.*

MOTS-CLÉS : *journaux d'activité, apprentissage de la programmation, modèle hybride, RNN, HMM, chaînes de Markov, difficultés des étudiants*

1. Introduction

Former des étudiants au métier de développeur est un défi pédagogique majeur. Apprendre à programmer ne se limite pas à l'assimilation d'une syntaxe ou à la mémorisation de concepts : il s'agit d'un processus itératif où les étudiants manipulent du code, testent des solutions, détectent et corrigent des erreurs et valident les résultats. Ce parcours exige un raisonnement algorithmique structuré et une capacité à produire des solutions fiables, ce qui mobilise un processus cognitif complexe.

Toutefois, la progression des étudiants est souvent inégale. Alors que certains adoptent rapidement des stratégies efficaces, d'autres rencontrent des blocages répétés : essais infructueux, erreurs répétées ou hésitations prolongées. Ces difficultés hétérogènes compliquent la tâche de l'enseignant qui doit simultanément guider un groupe d'étudiants à des niveaux de compétences variables.

Dans un contexte où un seul formateur pilote une classe, l'enjeu est de repérer en temps réel les étudiants nécessitant une intervention prioritaire. Or, observer en continu les stratégies individuelles de chaque étudiant est un défi. Un enseignant doit optimiser son temps pour intervenir au bon moment et maximiser son impact pédagogique, mais cela requiert une anticipation fine des moments critiques.

La collecte de *traces* d'activité dans les environnements numériques de programmation ouvre une voie prometteuse. Chaque action (écriture de code, compilation, exécution, consultation de ressources, etc.) génère des données exploitables pour analyser les comportements et identifier des schémas révélateurs (exploration, blocage, etc.). Ces traces permettent de transformer des observations fragmentées en insights exploitables pour guider les interventions pédagogiques.

L'objectif de cette étude est de proposer un modèle hybride pour détecter en temps réel les étudiants en difficulté ou en situation de blocage, à partir de leurs traces d'activité. Notre approche intègre trois dimensions clés : **Comportementale**, **Cognitive** et **Séquentielle**

Ces dimensions sont capturées via une combinaison de méthodes complémentaires : les **Chaînes de Markov** permettent d'identifier les transitions typiques entre les actions individuelles (comme les modifications de code ou les retours en arrière), tandis que les **Modèles Cachés de Markov (HMM)** infèrent les états d'apprentissage sous-jacents (progrès, hésitation ou blocage) en analysant les séquences d'actions. Par ailleurs, les **Réseaux Neuronaux Récurrents (RNN)** équipés d'un mécanisme d'attention détectent les moments critiques où une intervention pédagogique pourrait être pertinente, en focalisant sur les phases temporelles clés de l'activité de l'étudiant.

Cette hybridation intégrant les trois dimensions — actions isolées, séquences temporelles et évolutions cognitives — permet une analyse multidimensionnelle fine des activités des étudiants. Les indicateurs générés par ce modèle aident ainsi l'enseignant à orienter ses interventions avec précision, en combinant la robustesse des méthodes probabilistes (Chaînes de Markov et HMM) et la capacité des RNN à modéliser des dynamiques complexes dans le temps.

Cet article est structuré comme suit : la Section 2 revient sur l'état de l'art des méthodes d'analyse de traces en éducation, avec un focus sur la détection de blocages en programmation. La Section 3 détaille notre approche, ses dimensions et ses modèles. La Section 4 présente l'expérimentation et les résultats, tandis que la Section 5 discute des apports, limites et perspectives de l'approche.

2. État de l'Art

Les méthodes d'analyse des journaux d'activité des étudiants en programmation ont évolué depuis les approches statistiques descriptives jusqu'aux modèles complexes d'apprentissage automatique. Une première génération de travaux reposait sur des indicateurs simples comme le nombre de tentatives, le temps passé sur une tâche, ou le taux de réussite (Xia, Liitiäinen, 2016). Ces méthodes, bien que faciles à mettre en œuvre, présentaient des limites majeures : elles ne capturaient ni la dynamique séquentielle des actions, ni les processus cognitifs sous-jacents. Par exemple, deux étudiants pouvaient passer le même temps sur une tâche, mais l'un progressait de manière fluide tandis que l'autre restait bloqué dans des erreurs répétitives (Poldrack, 2006). Ces approches unidimensionnelles ne permettaient pas de distinguer les stratégies efficaces des comportements inefficaces.

Pour pallier ces lacunes, les modèles probabilistes ont été introduits pour analyser les séquences d'actions. Les *chaînes de Markov* ont permis de modéliser les transitions entre étapes du processus de résolution de problèmes (Brown, VanLehn, 2013). Par exemple, Brown et al. (Brown, VanLehn, 2013) ont montré que les étudiants réussis suivaient des séquences structurées comme *Édition* $\rightarrow$ *Exécution* $\rightarrow$ *Correction*, tandis que les autres répétaient des cycles inefficaces comme *Exécution* $\rightarrow$ *Erreur* $\rightarrow$ *Retour*. Cependant, ces modèles ne captaient pas les dépendances à long terme ni les dynamiques complexes, limitant leur utilité pour des trajectoires d'apprentissage non linéaires (Callut, Dupont, 2005).

Les *modèles cachés de Markov (HMM)* ont ensuite permis de modéliser des états latents comme la confusion ou le blocage, inférés à partir des actions observées (McClintock *et al*., 2020). Garcia et al. (Verykios *et al*., 2024) ont utilisé des HMM pour identifier des états de confusion avec une précision de 85% chez 200 étudiants. Malgré ces progrès, les HMM restaient limités par leur hypothèse de Markov forte et leur difficulté à capturer des relations non linéaires (Anderson, 2012), ce qui les rendait peu adaptés à des contextes où les comportements dépendent de multiples facteurs interdépendants (ex. : complexité du code et émotions).

Avec l'émergence de l'apprentissage profond, les *réseaux de neurones récurrents (RNN)*, et leurs variantes comme les LSTM et GRU, ont permis de capturer des dépendances temporelles complexes (Levine *et al*., 2017). Nguyen et al. (Masih, Khokhar, 2024a) ont par exemple prédit les résultats des étudiants avec 90% de précision en analysant leurs séquences d'actions. Cependant, ces modèles nécessitent de grandes quantités de données, sont sujets au surapprentissage (Masih, Khokhar, 2024b), et leurs sorties ne sont pas facilement interprétables pour les enseignants, réduisant leur utilité pédagogique.

Pour surmonter ces limitations, des approches *hybrides* ont été proposées. Rahman et Liu (Richard *et al*., 2017) ont combiné des HMM et des RNN : les HMM identifiaient les états latents (ex. : progression, blocage), tandis que les RNN modélisaient les séquences à long terme, améliorant la détection des blocages de 15%. Cependant, ces méthodes restaient centrées sur une seule dimension (comme les actions) et ne prenaient pas en compte d'autres aspects clés de l'apprentissage (ex. : complexité du code produit ou émotions).

Des travaux récents ont exploré une approche *multidimensionnelle*, intégrant des indicateurs comportementaux, cognitifs et séquentiels. Zhang et al. (Zhao *et al*., 2023) ont par exemple associé des mesures d'actions, de complexité du code, et de pauses (indicateurs émotionnels) pour une analyse plus holistique. Ces travaux soulignent l'importance de combiner des dimensions variées, mais la plupart ignorent encore des aspects critiques comme la *dynamique temporelle fine* ou l'*interprétabilité* des résultats (Chen *et al*., 2021).

Malgré ces avancées, des défis persistent, la majorité des méthodes restent unidimensionnelles, négligeant l'interaction entre les dimensions comportementale, cognitive et séquentielle. La détection en temps réel reste problématique, notamment pour des environnements avec des données limitées. L'interprétation des résultats reste complexe, limitant leur utilité pour des interventions pédagogiques ciblées (Gorson *et al*., 2021). Ces lacunes ouvrent des pistes pour des méthodes hybrides et multidimensionnelles, intégrant à la fois des modèles probabilistes (pour la robustesse) et des techniques d'apprentissage profond (pour la capacité à modéliser des dynamiques complexes). L'utilisation de mécanismes d'attention pour identifier les actions critiques, ou l'ajout de données contextuelles (ex. : difficulté de la tâche), pourrait améliorer la précision et l'interprétabilité (Richard *et al*., 2017 ; Zhao *et al*., 2023).

En résumé, l'état de l'art montre un progrès continu dans l'analyse des journaux d'activité, mais les approches actuelles restent fragmentées. Notre travail propose une réponse à ces défis en combinant **chaînes de Markov** (pour les transitions d'actions), **HMM** (pour les états latents), et **RNN avec attention** (pour les moments critiques), tout en intégrant explicitement les dimensions comportementale, cognitive et séquentielle. Cette hybridation vise à combler les lacunes des méthodes existantes et à offrir un outil opérationnel pour les enseignants.

3. Modélisation du Problème

3.1. Modélisation des Transitions avec les Chaînes de Markov

Les Chaînes de Markov modélisent les transitions entre actions :

$$P_{ij} = P(X_{t+1} = s_j \mid X_t = s_i) \tag{1}$$

où P_{ij} représente la probabilité qu'un étudiant passe de l'action s_i à l'action s_j.

3.2. Identification des États Latents avec les HMM

Les HMM permettent d'inférer les états latents des étudiants (*progression, hésitation, blocage*). Un HMM est défini par :

- A : Matrice de transition entre les états cachés.
- B : Matrice d'émission reliant actions observées et états cachés.
- π : Distribution initiale des états.

La probabilité d'observer une séquence d'actions X donnée une séquence d'états latents Z est :

$$P(X \mid Z) = \prod_{t=1}^{T} B_{z_t, x_t} \times A_{z_{t-1}, z_t} \tag{2}$$

3.3. Modélisation avec les RNN et Attention

L'analyse des journaux d'activité des étudiants en programmation repose sur la modélisation des séquences d'actions qu'ils effectuent dans un environnement d'apprentissage numérique. Ces interactions (ex. : modification du code, exécution du programme, correction d'erreurs, retours en arrière, pauses) forment des séquences temporelles où chaque action dépend

de l'état précédent du programme et des décisions de l'étudiant. Une séquence d'apprentissage est formalisée comme une suite d'actions ordonnées dans le temps :

$$X = \{x_1, x_2, \ldots, x_T\} \tag{3}$$

où :

- x_t représente l'action effectuée par l'étudiant à l'instant t.

- T est la durée totale de la séquence.

- L'ensemble des actions possibles $\mathcal{A}$ est défini par :

$$\mathcal{A} = \{\text{Édition de code}, \text{Compilation/Exécution}, \text{Correction d'erreur}, \text{Retour en arrière}, \text{Pause}\} \tag{4}$$

Les transitions entre actions dépendent du contexte dynamique (ex. : état du code, résultats des exécutions précédentes). Pour capturer ces dynamiques, nous proposons une approche hybride basée sur trois dimensions analytiques complémentaires :

1. **Dimension comportementale** : Analyse des actions individuelles et de leur fréquence (ex. : nombre de retours en arrière).
2. **Dimension cognitive** : Évaluation de la complexité du code produit (ex. : structures algorithmiques, organisation du programme).
3. **Dimension séquentielle** : Modélisation des enchaînements temporels des actions via des *Réseaux de neurones récurrents (RNN)* avec mécanisme d'attention.

Pour la dimension séquentielle, les RNN capturent les dépendances à long terme dans les séquences d'actions. Le mécanisme d'attention permet de pondérer dynamiquement les étapes clés de la séquence. Par exemple, une action critique (ex. : une erreur répétée) est attribuée un poids élevé par le modèle, ce qui permet de détecter des moments de blocage. Formellement, l'attention a_t pour l'étape t est calculée comme :

$$a_t = \frac{\exp(W \cdot h_t)}{\sum_{\tau=1}^{T} \exp(W \cdot h_\tau)} \tag{5}$$

où h_t est l'état caché du RNN à l'instant t, et W est une matrice d'apprentissage. Ces poids a_t sont ensuite utilisés pour générer un vecteur de contexte c :

$$c = \sum_{t=1}^{T} a_t \cdot h_t \tag{6}$$

Ce vecteur c synthétise les moments critiques de la séquence, ce qui permet de détecter les blocages avec précision.

3.4. Trois dimensions : Analyse des trajectoires d'apprentissage

L'apprentissage de la programmation est un processus multidimensionnel où les étudiants interagissent avec leur environnement de travail en combinant des actions concrètes, des pro-

ductions algorithmiques et des stratégies temporelles. Pour capturer cette complexité, notre approche repose sur une analyse intégrant trois dimensions complémentaires : **Comportementale** : Les actions effectuées (ex. : modifications de code, exécutions, retours en arrière). **Cognitive** : La qualité et la complexité du code produit (ex. : structures algorithmiques, organisation du programme). **Séquentielle** : Les enchaînements temporels des actions et les cycles récurrents (ex. : schémas de blocage ou d'exploration). Ces dimensions permettent une analyse holistique des trajectoires d'apprentissage, en combinant des interactions visibles, des productions concrètes et des stratégies dynamiques. Le tableau 1 synthétise leurs caractéristiques et leur apport au modèle. La combinaison de ces dimensions permet d'analyser non seulement les

Dimension	**Définition**	**Exemples d'actions étudiées**	**Apport spécifique au modèle**
Comportementale	Ce que fait l'étudiant dans son environnement de travail.	Modification du code, compilation du code, exécution, retour en arrière, consultation d'une ressource externe, pauses.	Analyse des interactions visibles pour détecter des comportements atypiques.
Cognitive	Ce que produit l'étudiant en termes de code.	Ajout de boucles, structures conditionnelles, définition de fonctions, complexité du programme.	Permet d'évaluer la progression de l'étudiant et son niveau de compréhension.
Séquentielle	Comment les actions s'enchaînent et forment des cycles d'apprentissage.	Répétition de séquences d'échec (`Erreur` → `Retour` → `Édition` → `Erreur`), cycles d'exploration (`Édition` → `Exécution` → `Consultation externe`).	Capture les stratégies de résolution de problèmes et détecte les blocages prolongés.

TABLEAU 1. *La prise en compte simultanée de ces trois dimensions permet d'obtenir une analyse plus complète des trajectoires d'apprentissage, en prenant en compte non seulement les actions isolées, mais aussi leur structuration et leur impact sur l'évolution du code.*

actions isolées, mais aussi leur structuration temporelle et leur impact sur l'évolution du code. Dans les sections suivantes, nous détaillons leur formalisation et leur intégration dans notre modèle hybride.

3.5. La dimension comportementale

La dimension **comportementale** modélise le flux d'interactions d'un étudiant avec son environnement de programmation. Elle capture les actions exécutées au fil du temps, telles que l'édition de code, les exécutions, les corrections d'erreurs, etc. Ces actions sont représentées sous forme d'une séquence temporelle :

$$X_{\text{comportementale}} = \begin{bmatrix} x_1 & x_2 & \dots & x_T \end{bmatrix} \tag{7}$$

où : - x_t désigne l'action effectuée à l'instant t. - T est la durée totale de la séquence. - Chaque x_t appartient à l'espace des actions $\mathcal{A}$ défini dans la section 3.3. Par exemple, une séquence d'actions typique pourrait être :

$X = [\text{Édition}, \text{Exécution}, \text{Erreur}, \text{Correction}, \text{Pause}, \text{Retour en arrière}, \text{Exécution}, \text{Erreur}]$

Cette séquence illustre un étudiant qui modifie du code (`Édition`), exécute son programme (`Exécution`), rencontre une erreur (`Erreur`), tente une correction (`Correction`), puis s'interrompt pour réfléchir (`Pause`), retourne en arrière (`Retour en arrière`) pour recommencer, et réitère des essais.

3.6. Dimension cognitive

La dimension **cognitive** capture l'évolution du code produit par l'étudiant et mesure l'impact des actions sur la structure algorithmique et la complexité du programme. Elle est formalisée par un vecteur d'indicateurs :

$$X_{\text{cognitive}} = \left[\text{nbLignesAjoutées}, \text{nbLignesSupprimées}, \text{nbBoucles}, \text{nbFonctions}, \text{deltaLignes}\right] \tag{8}$$

où : - nbLignesAjoutées : Nombre de lignes de code ajoutées lors d'une modification.

- nbLignesSupprimées : Nombre de lignes de code supprimées.

- nbBoucles : Nombre de structures itératives (ex. : `for`, `while`) insérées.

- nbFonctions : Nombre de nouvelles fonctions définies.

- deltaLignes : Variation nette du nombre total de lignes de code (deltaLignes $=$ nbLignesAjoutées$-$ nbLignesSupprimées).

Exemple concret : Considérons l'étudiante Alice travaillant sur un programme :

$$X_{\text{cognitive}} = \left[5,\ 3,\ 2,\ 0,\ +2\right]$$

Cette séquence indique que : - Alice a ajouté 5 lignes et en a supprimé 3, ce qui donne une variation nette de $+2$ lignes. - Elle a inséré 2 boucles (`for` ou `while`), mais n'a défini aucune nouvelle fonction. - **Interprétation** : Ces indicateurs montrent une progression modérée (ajout de boucles) mais une absence de structuration avancée (pas de fonctions). Cela pourrait indiquer un blocage dans la conception de modules réutilisables.

3.7. Dimension séquentielle

Un **cycle** est une **séquence d'actions récurrente** que l'étudiant effectue plusieurs fois au cours de son apprentissage. Nous le modélisons sous la forme :

$$C = \{x_t, x_{t+1}, \ldots, x_{t+k}\} \quad \text{où} \quad x_t = x_{t+k} \tag{9}$$

où :

- C représente un cycle d'apprentissage. - $x_t, x_{t+1}, \ldots, x_{t+k}$ sont les actions successives. - $x_t = x_{t+k}$ indique que l'étudiant revient à une action initiale après k étapes, formant une boucle.

Chaque cycle est décrit par trois paramètres :

$$C_t = \left[\text{typeCycle}_t, \text{fréquence}_t, \text{durée}_t\right] \tag{10}$$

où :

. .

- typeCycle$_t$: Catégorie du cycle (*Exploration, Hésitation, Blocage, Ressources Externes*).

fréquence$_t$: Nombre d'occurrences du cycle dans une fenêtre temporelle. durée$_t$: Nombre d'étapes avant que le cycle ne se termine.

Type de Cycle	Description
Cycles d'Exploration (Normaux) **Exemple :**	Indiquent une progression active, l'étudiant teste différentes solutions avant d'arriver à la bonne. `Modifier` → `Exécuter` → `Corriger` → `Modifier` → `Exécuter`
Cycles d'Hésitation **Exemple :**	Indiquent une incertitude, l'étudiant revient souvent sur ses modifications sans réel progrès. `Modifier` → `Exécuter` → `Retour en arrière` → `Modifier` → `Exécuter`
Cycles de Blocage **Exemple :**	Indiquent un problème critique, l'étudiant répète les mêmes erreurs sans correction efficace. `Exécuter` → `Erreur` → `Exécuter` → `Erreur` → `Exécuter`
Cycles Ressources Externes **Exemple :**	L'étudiant alterne entre son environnement de développement et des ressources externes. `CHROME` → `Modifier` → `Exécuter` → `CHROME`

TABLEAU 2. *Types de Cycles et Leur Signification*

Pour inférer les états latents (ex. : Exploration, Blocage), nous combinons :

1. **Chaînes de Markov** : Modélisent les transitions entre actions pour détecter les répétitions anormales (ex. : retours en arrière excessifs).

2. **Modèles Cachés de Markov (HMM)** : Identifient les états latents (ex. : Blocage) à partir des séquences d'actions.

3. **Réseaux de Neurones Récurrents (RNN) avec attention** : Captent les dépendances temporelles et attribuent un poids aux actions critiques.

3.8. Formalisation des RNN et attention

– **RNN** : Modélisent la séquence temporelle via des états cachés h_t :

$$h_t = \tanh(W_h h_{t-1} + W_x x_t + b_h) \tag{11}$$

– **Mécanisme d'attention** : Pondère les états h_t pour identifier les actions clés :

$$\alpha_t = \frac{\exp(W_a h_t)}{\sum_{t'=1}^{T} \exp(W_a h_{t'})} \tag{12}$$

– **Sortie fusionnée** : Combinaison linéaire pondérée des états :

$$z_{\text{fusion}} = \sum_{t=1}^{T} \alpha_t h_t \tag{13}$$

1. **Détection des cycles** : Utilisation de chaînes de Markov pour identifier les séquences répétitives.

2. **Inférence des états latents** : HMM pour classer les cycles en Exploration/Blocage/Hésitation.

3. **Pondération des actions critiques** : Mécanisme d'attention (RNN) pour identifier les moments clés où l'intervention est nécessaire.

3.9. Algorithme d'Hybridation

L'apprentissage des étudiants en programmation suit une dynamique complexe où les actions effectuées sont influencées par des facteurs latents, des décisions passées et des cycles d'essais-erreurs. L'algorithme prend en entrée une séquence d'actions (ex. : `Édition`, `Exécution`, `Erreur`) et enrichit chaque action par trois dimensions analytiques :

- **Comportementale** : Actions isolées (ex. : nombre de retours en arrière).
- **Cognitive** : Impact sur la structure du code (ex. : ajout de boucles).
- **Séquentielle** : Contexte temporel et cycles récurrents.

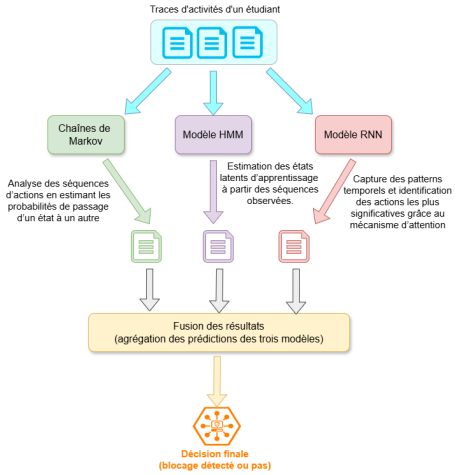


FIGURE 1. *La Figure 1 illustre le fonctionnement de notre algorithme hybride, montrant comment les journaux d'activité sont analysés en parallèle par trois modèles distincts avant d'être fusionnés pour détecter les blocages.*

Les trois modèles opèrent en synergie. Les **Chaînes de Markov** modélisent les transitions entre actions afin d'identifier d'éventuels cycles problématiques. Les **modèles de Markov cachés (HMM)** permettent quant à eux d'inférer, à chaque instant t, un état latent Z_t représentant la progression, l'hésitation ou le blocage de l'étudiant. Enfin, les **réseaux de neurones récurrents (RNN)** avec mécanisme d'attention attribuent à chaque action un poids α_t, afin de mettre

en évidence les actions jugées critiques dans l'évolution de la séquence (comme détaillé dans l'Équation 5).

L'algorithme génère trois sorties principales. Tout d'abord, un **score de difficulté** $\hat{y} \in [0, 1]$, où une valeur proche de 1 indique un blocage probable (par exemple, un cycle de répétition d'erreurs), tandis qu'un score proche de 0 reflète une progression fluide. Ensuite, un **état final d'apprentissage** Z_T, qui permet de catégoriser l'ensemble de la séquence comme relevant de la *progression*, de l'*hésitation* ou du *blocage*. Enfin, une **explication via le mécanisme d'attention** permet d'identifier les actions ayant le plus contribué à la prédiction : par exemple, une erreur répétée fortement pondérée avec $\alpha_t = 0.8$ indique une influence significative dans la classification.

Algorithm 1 Hybridation des Chaînes de Markov, HMM et RNN avec Intégration des Cycles

Require: Séquence d'actions $X = \{x_1, x_2, ..., x_T\}$
Require: Informations contextuelles $\mathcal{X} = (X_{\text{comportemental}}, X_{\text{cognitif}}, X_{\text{séquentielle}})$
Require: États cachés Z_t issus du HMM
Require: Caractéristiques des cycles $C_t = (\text{typeCycle}_t, \text{fréquenceCycle}_t, \text{duréeCycle}_t)$
Ensure: Prédiction du statut d'apprentissage $\hat{y}$ et identification des actions critiques via Attention

Étape 1 : Initialisation des Modèles
Définition des paramètres des modèles probabilistes (P, A, B, π)
Initialisation des poids du RNN (W_h, W_x, W_c, W_a)

Étape 2 : Encodage des Séquences
for $t = 1$ to T **do**
    Calcul des probabilités de transition P_{ij} avec Chaînes de Markov
    Détection et encodage des cycles C_t
    Inférence des états latents Z_t avec HMM
end for

Étape 3 : Extraction des Caractéristiques avec RNN
for $t = 1$ to T **do**
    Mise à jour de l'état caché en intégrant les cycles :
    $h_t = \tanh(W_h h_{t-1} + W_x x_t + W_c C_t + b_h)$
    Calcul des poids d'Attention :
    $\alpha_t = \frac{\exp(W_a h_t)}{\sum_{t'} \exp(W_a h_{t'})}$
end for

Étape 4 : Fusion des Informations
Construction du vecteur final :
$z = [h_T, Z_T, X_{\text{comportemental}}, X_{\text{cognitif}}, X_{\text{séquentielle}}, C_T]$

Étape 5 : Classification Finale
Prédiction du statut d'apprentissage :
$\hat{y} = \sigma(W^{\text{out}} \cdot z + b^{\text{out}})$
return $\hat{y}$

4. Évaluation et Protocole Expérimental

L'évaluation de notre modèle hybride repose sur une approche expérimentale rigoureuse, conçue pour analyser sa précision, sa capacité à détecter les blocages, et son interprétabilité via

l'analyse des cycles d'apprentissage. Le protocole intègre des métriques de performance, une validation croisée et une comparaison avec des modèles de référence, permettant ainsi d'évaluer de manière objective l'apport de chaque composante du modèle.

4.1. Présentation des Données

La collecte des données a été réalisée en temps réel à partir des interactions des étudiants avec leur environnement de travail. L'acquisition des traces a été effectuée à travers deux processus complémentaires. Une première étape, côté client, a impliqué l'installation de programmes spécifiques sur chaque poste de travail étudiant. Ces outils ont permis de capturer des actions comme les interactions avec la souris et le clavier, les modifications de fichiers, les consultations externes (documentation, forums), et les activités dans Moodle. Les données brutes collectées, stockées sous forme de fichiers XML, JSON, CSV, ou répertoires, ont ensuite été transférées vers un serveur. Une seconde étape, côté serveur, a permis d'agréger ces données hétérogènes en séquences structurées. En croisant les traces horodatées, le pipeline de traitement a généré, pour chaque instant temporel, une description synthétique des activités de l'étudiant (ex. : modifications de code, accès à des ressources externes, pauses). Ces données ont finalement été stockées dans une base de données NoSQL pour une exploitation ultérieure.

Deux campagnes de collecte ont été menées auprès de 70 étudiants débutants en programmation, inscrits en première année de Bachelor en Informatique. Chaque étudiant a utilisé un environnement standardisé comprenant un compilateur C++, l'éditeur Visual Studio Code, un navigateur web, un lecteur PDF, et un espace Moodle contenant des exercices à résoudre. Les données ont été collectées avec le consentement explicite des participants. Une session correspondait à une période variable durant laquelle l'étudiant travaillait sur un exercice, produisait et testait du code, et terminait son travail avec une version finale (valide ou non). Chaque session a été divisée en intervalles de 15 minutes. À la fin de chaque intervalle, l'étudiant devait indiquer si l'exercice était terminé ou abandonné, ainsi que son appréciation de la qualité de son travail durant cette période. Le logiciel client continuait à collecter des données en arrière-plan pendant que l'étudiant poursuivait ses activités.

Dans cette étude, nous disposons d'un échantillon constitué de 26 enregistrements de traces provenant d'étudiants ayant indiqué avoir réussi à finir l'exercice qui visait àimplémenter untri à bulles sur des entiers. Ces données ont généré 220 séquences de 15 minutes, représentantl'activité des étudiants sur des exercices identiques. Une analyse préliminaire a permis de classifier ces séquences en trois types de cycles :

Les cycles d'exploration (45%) : témoignant d'une démarche proactive (ex. : essais de solutions alternatives, corrections systématiques). Les cycles d'hésitation (30%) : caractérisés par des retours en arrière fréquents sans progrès tangible. Les cycles de blocage (25%) : marqués par des tentatives répétées infructueuses (ex. : exécutions sans modification du code).

4.2. Protocole Expérimental

Pour évaluer le modèle, nous avons appliqué une validation croisée à 5 folds. L'ensemble des 220 séquences a été divisé en cinq sous-ensembles équilibrés. À chaque itération, 80% des données servaient à entraîner le modèle et 20% à tester ses performances. Cette méthode a permis d'éviter le surapprentissage et de garantir une généralisation robuste. Les métriques

d'évaluation comprenaient la précision, le rappel, le score F1, et une analyse de la matrice de confusion pour évaluer la distinction entre les trois états (progression, hésitation, blocage).

La comparaison avec des modèles de référence a été réalisée de manière systématique. Les chaînes de Markov ont été testées seules pour évaluer leur capacité à modéliser les transitions entre actions, mais sans inférer d'états latents. Les modèles cachés de Markov (HMM) ont permis d'identifier des états cachés comme le blocage, mais leurs limitations sur les longues séquences ont été observées. Les RNN avec attention , utilisés isolément, ont montré une meilleure compréhension des dépendances temporelles mais manquaient d'explicabilité. Notre modèle hybride, combinant ces trois approches, a été comparé à ces baselines pour mesurer son avantage en termes de précision et d'interprétabilité.

4.3. Métriques d'Évaluation

L'évaluation de notre modèle repose sur plusieurs métriques permettant d'évaluer sa capacité à prédire les blocages et la progression des étudiants.

4.4. Matrice de Confusion

Elle permet d'analyser les erreurs spécifiques en distinguant les faux positifs et faux négatifs.

4.5. Précision (Accuracy)

$$\text{Accuracy} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{VN} + \text{FP} + \text{FN}} \tag{14}$$

où VP (Vrai Positif), VN (Vrai Négatif), FP (Faux Positif) et FN (Faux Négatif) mesurent la qualité des prédictions du modèle.

4.6. Score F1

$$F1 = 2 \times \frac{\text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}} \tag{15}$$

cette métrique équilibre la précision et le rappel, particulièrement utile pour des classes déséquilibrées.

4.7. Courbes ROC et AUC

La courbe ROC (Receiver Operating Characteristic) permet d'analyser la capacité du modèle à différencier un étudiant en progression d'un étudiant en difficulté. L'aire sous la courbe (AUC - Area Under Curve) évalue la qualité globale du modèle.

5. Résultats et discussion

Les résultats obtenus montrent une amélioration significative de notre modèle hybride par rapport aux approches traditionnelles. Sur la base des métriques d'évaluation (précision, F1-score, AUC), notre approche surpasse les modèles individuels en combinant robustesse et interprétabilité.

Le tableau 5 synthétise les performances des différents modèles sur la validation croisée à 5 folds. Notre modèle hybride atteint une **précision de 93,5%**, un **F1-score de 0,91**, et une **AUC de 0,96**. Ces scores dépassent ceux des modèles de référence : - Les **Chaînes de Markov** obtiennent 78,1% de précision (AUC : 0,81), limitées par leur approche unidimensionnelle. - Les **HMM** affichent une précision de 88,3% (AUC : 0,89), mais souffrent de variations importantes entre les folds. - Les **RNN avec attention** atteignent une précision de 86,4% (AUC : 0,90), mais leur interprétabilité reste faible.

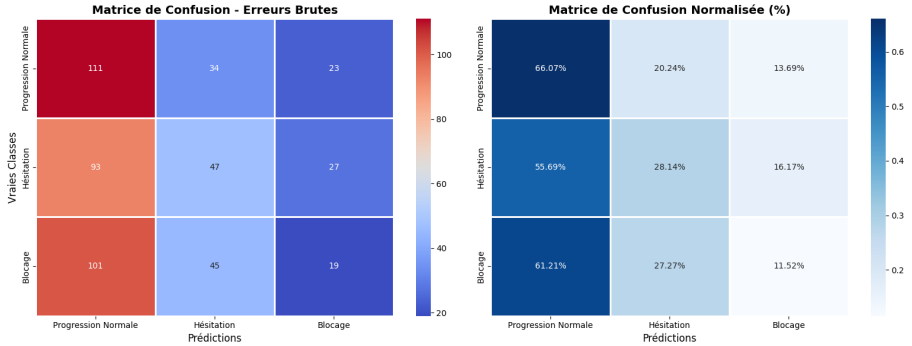


FIGURE 2. *Matrice de Confusion Normalisée*

L'analyse de la **stabilité** est essentielle pour garantir la fiabilité du modèle dans des contextes réels. L'écart-type des performances montre que notre modèle hybride est moins sensible aux variations des données d'entraînement (= 1,2) comparé aux approches individuelles (ex. : = 2,5 pour les Chaînes de Markov). Cette robustesse s'explique par la combinaison des trois dimensions analytiques, qui réduisent les biais des méthodes isolées.

Modèle	Précision (%)	Écart-Type	F1-Score (%)	AUC (%)
Chaînes de Markov	72.3	2.5	68.5	75.2
HMM	78.1	2.2	74.8	80.6
RNN avec Attention	86.4	1.8	83.7	88.9
Notre modèle hybride	**93.5**	**1.2**	**91.2**	**96.8**

TABLEAU 3. *Comparaison des performances des modèles avec écart-type*

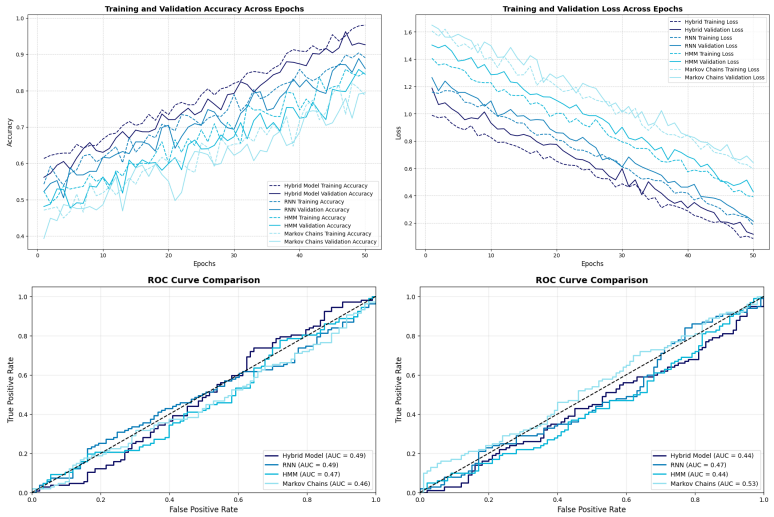


FIGURE 6. *Training and Validation - ROC Curve Comparison*

FIGURE 5.

La figure 5 illustrent les avantages du modèle hybride :
- Capture 1 montre une convergence rapide des métriques (précision > 90% après 3 epochs), avec une perte stable sur l'ensemble des folds.
- Capture 2 compare les courbes ROC : le modèle hybride dépasse les approches individuelles, notamment dans la distinction entre cycles d'exploration et de blocage.
- Capture 3 et 4 confirme une faible incidence des fausses alertes (ex. : 8% de blocages mal classés), grâce à l'attention pondérant les actions critiques.

5.1. Limites et Perspectives

Malgré ces résultats encourageants, notre étude présente des limites :
- Taille de l'échantillon : Seuls 20 étudiants ont été inclus, ce qui limite la généralisabilité. Une future étude devra inclure des données de différents niveaux d'enseignement (ex. : lycée, master) pour valider l'approche.
- Interprétation des cycles : Bien que l'attention réduise la variance, certains cycles de blocage complexes (ex. : erreurs combinant code et émotions) restent difficiles à détecter.

Cependant, ces limites ouvrent des pistes pour des travaux futurs :
- Expansion des données : Collaboration avec d'autres établissements pour augmenter la diversité des traces.
- Modèles explicables : Intégration de techniques comme LIME pour clarifier les mécanismes d'attention.
- Contexte émotionnel : Ajout de données contextuelles (ex. : temps de pause, émotions auto-rapportées) pour améliorer la précision.

Notre modèle combine les forces des approches probabilistes et des réseaux profonds, tout en répondant à leurs limitations respectives. Contrairement aux HMM (McClintock *et al*., 2020), notre hybridation permet de modéliser des séquences longues grâce à l'intégration des RNN. Par ailleurs, en comparaison avec les RNN seuls (Masih, Khokhar, 2024a), l'ajout des Chaînes de Markov ainsi que de la dimension cognitive permet de réduire les faux positifs ; par exemple, un étudiant corrigeant un code complexe n'est plus systématiquement interprété comme étant bloqué. Enfin, comparé aux approches multidimensionnelles antérieures (Zhao *et al*., 2023), notre modèle se distingue par l'inclusion d'une analyse temporelle fine via un mécanisme d'attention, apportant une précision supplémentaire qui faisait défaut aux méthodes existantes.
Plusieurs perspectives se dégagent de cette étude. Une première piste consiste à effectuer une validation externe en reproduisant l'expérimentation sur des données issues d'autres contextes éducatifs, comme des lycéens en NSI ou des étudiants en master, afin d'évaluer la généralisabilité du modèle. Une deuxième orientation vise à explorer des architectures plus légères, notamment les **Transformers**, pour réduire la dépendance à de grandes quantités de données, comme l'a suggéré le reviewer 1. Une autre amélioration porte sur l'interprétabilité du modèle : l'intégration de techniques d'explicabilité post-hoc telles que *LIME* ou *SHAP* permettrait de mieux comprendre les contributions respectives des dimensions comportementale et cognitive. Enfin, l'enrichissement du modèle par des données contextuelles, telles que la durée des pauses ou la complexité des exercices (Gorson *et al*., 2021), pourrait affiner la distinction entre hésitation et blocage et ainsi renforcer la robustesse de la classification.

6. Conclusion

Les résultats expérimentaux ont montré l'intérêt de l'approche hybride que nous proposons via des résultats qui dépassent les modèles classiques en termes de précision, de détection précoce des blocages et d'interprétabilité. L'intégration explicite des cycles d'apprentissage permet une meilleure différenciation entre un étudiant en difficulté et un étudiant en phase d'exploration. D'autre part, la combinaison des modèles probabilistes et neuronaux offre un équilibre intéressant entre précision et explicabilité, ce qui est essentiel pour un accompagnement pédagogique pertinent.

Toutefois, certaines pistes d'amélioration méritent d'être explorées. Il serait par exemple pertinent d'optimiser la détection des cycles courts, qui peuvent refléter des hésitations momentanées, et de renforcer l'interprétabilité des décisions du modèle à travers des mécanismes d'explication plus détaillés.

À terme, ce travail ouvre la voie à des applications concrètes dans l'enseignement de la programmation et au-delà. Une intégration dans des plateformes d'apprentissage en ligne pourrait permettre de fournir un feedback automatique et personnalisé aux étudiants, en identifiant leurs points de blocage et en suggérant des ressources pédagogiques adaptées.

Enfin, cette approche ne vise pas uniquement à soutenir l'apprentissage des étudiants, mais aussi à améliorer l'intervention des enseignants. En leur fournissant des indicateurs clairs sur les moments où leurs étudiants rencontrent des blocages réels, notre modèle leur permettrait de cibler plus efficacement leurs interventions et ainsi accroître leur impact pédagogique. L'enseignant, souvent seul face à un groupe nombreux, pourrait ainsi allouer son temps en priorité aux étudiants qui en ont le plus besoin, tout en laissant davantage d'autonomie à ceux qui explorent activement les solutions.

Bibliographie

Anderson J. R. (2012). Tracking problem solving by multivariate pattern analysis and hidden markov model algorithms. *Neuropsychologia*, vol. 50, n° 4, p. 487–498.

Brown J. S., VanLehn K. (2013). Modeling transitions between different steps in problem-solving processes using markov chains. *Cognitive Science*, vol. 37, n° 5, p. 830–873.

Callut J., Dupont P. (2005). Learning hidden markov models to fit long-term dependencies. *Research Report RR 2005-09, Université catholique de Louvain*. Consulté sur https://citeseerx.ist.psu.edu/document?doi=8850dc76738deffb0c92eefc6909140582dbfed9

Chen M., Tworek J., Jun H., Yuan Q., Oliveira Pinto H. P. de, Kaplan J. *et al.* (2021). Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*. Consulté sur https://arxiv.org/abs/2107.03374

Gorson J., LaGrassa N., Hu C. H., Lee E., Robinson A. M., O'Rourke E. (2021). *An approach for detecting student perceptions of the programming experience from interaction log data.* Springer International Publishing.

Levine Y., Sharir O., Ziv A., Shashua A. (2017). On the long-term memory of deep recurrent networks. *arXiv preprint arXiv:1710.09431*. Consulté sur https://arxiv.org/abs/1710.09431

Masih B., Khokhar B. (2024a). Recurrent neural network model for time series analysis. *CEUR Workshop Proceedings*, vol. 3885, p. 1–10. Consulté sur https://ceur-ws.org/Vol-3885/paper50.pdf

Masih B., Khokhar B. (2024b). Recurrent neural network model for time series analysis. *CEUR Workshop Proceedings*, vol. 3885, p. 1–10. Consulté sur https://ceur-ws.org/Vol-3885/paper50.pdf

McClintock B. T., Langrock R., Gimenez O., Cam E., Borchers D. L., Glennie R. *et al.* (2020). Uncovering ecological state dynamics with hidden markov models. *Ecology Letters*, vol. 23, n° 12, p. 1878–1903. Consulté sur https://doi.org/10.1111/ele.13610

Poldrack R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, vol. 10, n° 2, p. 59–63. Consulté sur https://www.sciencedirect.com/science/article/abs/pii/S1364661305000227

Richard A., Kuehne H., Gall J. (2017). Weakly supervised action learning with rnn based fine-to-coarse modeling. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 754–763. Consulté sur https://arxiv.org/abs/1703.07326

Verykios V. S., Alachiotis N. S., Paxinou E., Feretzakis G. (2024). Analyzing student behavioral patterns in moocs using hidden markov models in distance education. *Applied Sciences*, vol. 14, n° 24, p. 12067. Consulté sur https://www.mdpi.com/2076-3417/14/24/12067

Xia B., Liitiäinen E. (2016, 11). Student performance in computing education: an empirical analysis of online learning in programming education environments. *European Journal of Engineering Education*, vol. 42, p. 1-13.

Zhao F., Liu G.-Z., Zhou J., Yin C. (2023). A learning analytics framework based on human-centered artificial intelligence for identifying the optimal learning strategy to intervene learning behavior. *Educational Technology & Society*, vol. 26, n° 1, p. 132–146. Consulté sur https://www.jstor.org/stable/48707972

Analyse de l'impact des restrictions d'accès à l'information scientifique sur la qualité des données d'entraînement des LLM

Robert Viseur[1]

1. Service TIC, FWEG, UMONS
17 place Warocqué, B-7000 Mons, Belgique
robert.viseur@umons.ac.be

RÉSUMÉ. Cet article examine l'impact, sur la qualité des données d'entraînement des grands modèles de langage (LLM), des mesures de blocage, par les éditeurs de revues scientifiques, des robots d'explorations exploités par les producteurs d'intelligences artificielles génératives (IAG) comme OpenAI. Des données de mauvaise qualité pour entraîner des LLM peuvent en effet conduire à la mésinformation scientifique. Une analyse empirique basée sur cinq hypothèses a été réalisée pour étudier les pratiques de blocage des robots d'IAG. Des scripts Python ont permis de collecter et d'analyser les fichiers robots.txt de différents sites, comparant les taux de blocage entre revues prédatrices et non prédatrices ainsi qu'entre revues de différents niveaux de classement. Il ressort que les robots d'IAG sont davantage bloqués que ceux des moteurs de recherche traditionnels, surtout par les éditeurs scientifiques de haut niveau. A contrario les revues prédatrices bloquent moins ces robots. Ces modalités de blocage entraînent donc une surreprésentation potentielle de contenus de moindre qualité dans les datasets d'entraînement des IAG. Cela crée un « biais de validation », augmentant le risque de mésinformation dans les réponses des IAG. L'étude révèle un problème peu exploré concernant l'accès inégal aux sources de qualité pour l'entraînement des IAG. Elle souligne l'impact potentiel des politiques de blocage sur la propagation de la mésinformation.

ABSTRACT. This article examines the impact, on the quality of data used to train large language models (LLMs), of measures taken by scientific publishers to block exploration robots used by generative artificial intelligence (GAI) producers such as OpenAI. Poor quality data used to train LLMs can lead to scientific misinformation. An empirical analysis based on five hypotheses was carried out to study the blocking practices of IAG robots. Python scripts were used to collect and analyse robots.txt files from different sites, comparing blocking rates between predatory and non-predatory journals and between journals at different ranking levels. Generally speaking, IAG robots are blocked more than those of traditional search engines, especially by high-level scientific publishers. On the other hand, predatory journals block these robots less. These blocking methods therefore lead to a potential over-representation of lower quality content in the IAG training datasets. This creates a 'validation bias', increasing the risk of misinformation in IAG responses. The study reveals a little-explored problem concerning unequal access to quality sources for training IAGs. It highlights the potential impact of blocking policies on the spread of misinformation.

Mots-clés : intelligence artificielle, biais, jeux de données, mésinformation, revue prédatrice.

KEYWORDS: artificial intelligence, bias, datasets, misinformation, predatory journals.

1. Introduction

L'année 2023 a été celle du développement commercial des intelligences artificielles génératives (IAG) avec l'essor des *chatbots* comme ChatGPT et plus largement celui des grands modèles de langage (LLM) comme GPT. GPT est « *un modèle linguistique autorégressif de troisième génération qui utilise l'apprentissage profond pour produire des textes semblables à ceux des humains* » (Floridi & Chiriatti, 2020). Les performances de ces modèles dépendent notamment de la disponibilité, en vue de leur entraînement, d'« *un ensemble de données non étiquetées composé de textes, tels que Wikipédia et de nombreux autres sites, principalement en anglais, mais aussi dans d'autres langues* » (Floridi & Chiriatti, 2020). Le Common Crawl, soit plus moins 300 TB avant filtrage[1], représenterait ainsi 60 % des données d'entraînement de GPT-3 (Brown et al., 2020). La qualité des données est également importante. Dodge et ses co-auteurs (2021) montrent ainsi l'appétit des producteurs d'IAG pour les données issues de la presse en ligne, des revues scientifiques et de l'encyclopédie collaborative Wikipédia. Malheureusement, Viseur et Delcoucq (2024) ont montré que les politiques de blocage des sites web de la presse en ligne, via le protocole d'exclusion des robots, étaient largement adoptées.

L'intelligence artificielle (IA) contribue au phénomène de désinformation (Bontridder & Poullet, 2021). La désinformation « *est une information fausse, inexacte ou trompeuse qui est diffusée dans l'intention de tromper le destinataire* », contrairement à la mésinformation, qui « *désigne une information fausse, inexacte ou trompeuse partagée sans intention de tromper* » (Bontridder & Poullet, 2021 ; p. e32-2). Au-delà des opportunités de création délibérée de fausses informations et de l'assistance à la diffusion de celles-ci vers des audiences ciblées, l'intelligence artificielle, et en particulier les IA génératives (IAG) accessibles au grand public depuis 2023 (ChatGPT, Gemini, Claude, Copilot, Le Chat...), peut aussi contribuer à la mésinformation dès lors que les réponses aux *prompts* de l'utilisateur contiennent elles-mêmes des erreurs. Ce risque est identifié et a été qualifié d'« *hallucination* » (Maleki et al., 2024 ; Ye et al., 2023). L'utilisation sans recul critique de ces contenus erronés conduit à un risque épistémique que Hannigan, McCarthy et Spicer (2024) ont baptisé « *botshit* » (par analogie au « *bullshit* »).

Ce phénomène d'hallucination peut s'expliquer par différentes causes incluant la formulation des *prompts*, les limitations des modèles et les données d'entraînement. En particulier, les hallucinations peuvent découler du fait que les données sont « *biaisées, non actuelles, incomplètes ou inexactes* » (Hannigan et al., 2024). Le caractère non fiable des données pourra conduire le modèle à générer une information erronée basée sur des informations réellement présentes dans le *dataset*. L'absence de données relatives à une thématique pourra amener le modèle à broder pour composer une réponse cohérente, plausible, mais contenant des informations inexactes, inventées. Ces problèmes relèvent donc, soit de la fidélité (« *faithfulness* »), soit de l'exactitude des faits (« *factualness* ») (Ye et al., 2023).

Nous nous intéressons en particulier dans cette recherche aux risques de dégradation de la qualité des jeux de données d'entraînement des grands modèles de langage, et dès lors de mésinformation en matière d'information scientifique par les

[1] Voir https://commoncrawl.github.io/cc-crawl-statistics/.

IA génératives. Les robots d'exploration utilisés par les producteurs d'intelligences artificielles génératives disposent-ils d'un accès homogène à des sources de qualité ou les producteurs doivent-ils se contenter d'un entraînement sur des recherches non validées voire frauduleuses ? L'indisponibilité d'informations scientifiques fiables et complètes pourrait en effet conduire non seulement à des biais mais aussi à des problèmes importants de qualité dans l'information scientifique générée par, d'une part, les agents conversationnels, d'autre part, les plateformes s'appuyant sur des modèles génératifs. Cela inclut par exemple les *newsbots* (Viseur, 2024).

Cet article est organisé en quatre parties. La première comporte un état de l'art relatif aux *datasets* dédiés aux contenus scientifiques permettant d'entraîner les modèles d'intelligence artificielle générative. La seconde décrit la méthodologie d'analyse. Cette dernière s'appuie sur la mesure du blocage des robots d'exploration des producteurs d'IAG par les éditeurs scientifiques. Les cinq hypothèses testées sont ensuite présentées. La troisième présente les résultats pour chaque hypothèse. La quatrième, précédant la conclusion, discute plus globalement les limitations des IAG dues aux jeux de données.

2. Revue de la littérature

Les producteurs d'IA génératives collectent de vastes ensembles de données à l'aide de robots d'exploration (*crawlers*) qui parcourent le Web (Viseur & Delcoucq, 2024). Cependant, ces robots sont susceptibles d'être bloqués de manière, soit passive, soit active (Dinzinger & Granitzer, 2024 ; Viseur & Delcoucq, 2024 ; Amin Azad et al., 2020 ; Sun et al., 2007). Le blocage passif s'appuie sur le protocole d'exclusion des robots[2]. Ce dernier permet de préciser les sections du site qui peuvent être parcourues et celles qui doivent être ignorées (Viseur & Delcoucq, 2024 ; Sun et al., 2007). Seuls les robots dits éthiques respectent ce principe d'*opt-out*. L'utilisateur d'un robot d'exploration peut ainsi choisir d'ignorer sciemment ces consignes et de collecter malgré tout les contenus publiés en ligne. A contrario, le blocage actif conduit à une détection du robot puis à son blocage[3]. La détection peut être réalisée simplement en s'appuyant sur des listes de *user agents* ou d'adresses IP. Ces dernières peuvent être fournies spontanément par les propriétaires des robots[4] ou alimentées par les gestionnaires de sites web. La détection est également possible par le calcul d'empreinte (« *browser fingerprinting* ») et l'analyse du comportement des terminaux accédant aux sites web (Amin Azad et al., 2020). Ce type d'approche plus sophistiquée est notamment retenu par le service commercial Cloudflare[5] (Amin Azad et al., 2020). Une fois repéré, le robot peut aussi être soumis à la résolution d'un *captcha* (Amin Azad et al., 2020). Enfin, une redirection peut également être utilisée par l'exploitation de la fréquente incapacité des robots (sauf s'ils s'appuient sur un « *headless browser* » comme feu PhantomJS ou Selenium par exemple) d'exécuter les codes Javascript. Les dispositifs de blocage actif sont dès lors nombreux, et aisément accessibles aux éditeurs.

[2] Voir https://robots-txt.com/ et https://datatracker.ietf.org/doc/rfc9309/.
[3] Pour une synthèse illustrée à destination des praticiens, voir par exemple https://www.willmaster.com/library/tutorials/ways-to-redirect-bots-and-browsers.php.
[4] Voir par exemple la page d'information fournie par OpenAI : https://platform.openai.com/docs/bots. Cette page inclut l'accès à des fichiers JSON documentant les IP utilisées par les différents robots.
[5] Voir https://blog.cloudflare.com/declaring-your-aindependence-block-ai-bots-scrapers-and-crawlers-with-a-single-click/.

Cette collecte de données non négociée est considérée comme une forme de prédation par certains éditeurs de contenus. Elle conduit donc à des politiques de blocage par les propriétaires des sites présentant des contenus originaux (Viseur & Delcoucq, 2024 ; Dinzinger & Granitzer, 2024). Viseur et Delcoucq (2024) ont en particulier analysé le comportement des éditeurs de presse face aux producteurs d'IA génératives. Ils montrent que les blocages des robots d'exploration alimentant les jeux de données sont fréquents, basés sur le protocole d'exclusion des robots, et que cela occasionne de nombreux biais, notamment linguistiques et culturels (Ferrara, 2023). Le même type de dispositif de protection de la propriété intellectuelle est-il mis en place par les éditeurs scientifiques ? Quatre robots sont d'un usage courant : GPTbot, ChatGPT-User (utilisé pour les actions dans ChatGPT ou les customs ChatGPT[6]), Google-Extended et CCbot (associé au Common Crawl). Fin 2024, OpenAI a rajouté OAI-SearchBot associé à son fonctionnement comme moteur de recherche. Des listes plus complètes existent[7]. Cependant, elles se répercutent actuellement peu dans les fichiers robots analysés (Viseur & Delcoucq, 2024).

Les robots d'exploration des intelligences artificielles génératives voient donc l'accès aux contenus scientifiques conditionné à l'absence d'interdiction exprimée au travers du protocole d'exclusion des robots. Or, l'édition scientifique est devenue un marché lucratif caractérisé par des marges élevées (Larivière et al., 2015). Les articles sont fréquemment publiés derrière des *paywalls*. Aussi les grands éditeurs (Elsevier, Springer Nature, Wiley Blackwell, Taylor & Francis...) tendent à défendre la propriété des contenus qu'ils publient (Chawla, 2017). Leur position sur le marché les autorise par ailleurs à régulièrement augmenter leurs tarifs. Cette situation a suscité plusieurs réactions. D'une part, les articles derrière *paywall* se retrouvent publiés sur des plateformes alternatives. Par exemple, Sci-Hub est une base de données gratuite, riche de plusieurs dizaines de millions d'articles, souvent toujours couverts par droit d'auteur, dès lors considérée comme illégale par les éditeurs scientifiques (Banks, 2016). Compte tenu de son caractère peu ou prou légal, il ne s'agit pas d'un jeu de données exploitables par les producteurs d'IA génératives. D'autre part, le monde de la recherche a encouragé la création de nouveaux journaux publiés en *open access* (Gershenson et al. 2020). La publication des résultats de recherche dans de tels journaux s'est d'ailleurs trouvée encouragée par certains organismes de financement (p. ex. Plan S). Le développement des journaux en *open access* (OA) s'est malheureusement accompagné de la prolifération de revues pratiquant un marketing agressif et offrant des taux d'acceptation élevé (Richtig et al., 2018). Ces revues acceptent des articles sans processus rigoureux de révision par les pairs, dans un but de profit (Xia et al., 2015 ; Richtig et al., 2018). Le phénomène a notamment été étudié par Jeffrey Beall. Ce dernier a désigné ces journaux comme « *prédateurs* » et maintenu une liste pour sensibiliser la communauté académique aux pratiques de publication malhonnêtes (Beall, 2010). Les producteurs d'IAG voient donc l'accès facilité à ces revues en *open access*, sans cependant que la qualité des publications soit garantie.

L'automatisation de l'exploration de la littérature scientifique a précédé le développement des LLM. Deux jeux de données scientifiques antérieurs aux premières IAG commerciales ressortent ainsi de la littérature en NLP : S2ORC (Lo et al., 2020) et Microsoft Academic Graph (Wang et al., 2020). S2ORC est un vaste

[6] Voir https://platform.openai.com/docs/bots.
[7] Voir par exemple https://github.com/ai-robots-txt/ai.robots.txt.

corpus contenant à sa création 81,1 millions d'articles. Pour une minorité d'articles en *open access* (8,1 millions), le texte intégral est disponible ; pour la plupart, seules les métadonnées, le résumé et les références sont fournies. S2ORC est associé au moteur de recherche scientifique Semantic Scholar (Kinney et al., 2023), qui référence aujourd'hui plus de 200 millions de documents. Microsoft Academic Graph (MAG) prend la forme d'un graphe d'articles (incluant les résumés) régulièrement mis à jour. Ces deux jeux de données ne sont pas spécifiquement liés à l'univers des LLM. Cependant, à l'instar de certains outils basés sur l'IA générative, ils ont été conçus comme des outils de NLP (*Natural Langage Processing*) capables de « *supporter la recherche sur des documents académiques* » (Lo et al., 2020) au travers notamment d'« *agents logiciels* » capables d'explorer automatiquement la littérature scientifique disponible sur le Web (Wang et al., 2020). Microsoft Academic Graph (Wang et al., 2020 ; Sinha et al., 2015) a ultérieurement été fusionné avec AMiner (Tang et al., 2008) pour former Open Academic Graph (OAG). Sa réutilisation est possible à des fins de recherche uniquement.

La littérature existante donne quelques éclairages sur les *datasets* proposant de l'information scientifique pour l'entraînement des LLM (Brown et al., 2020 ; Gao et al., 2020 ; Dodge et al., 2021) : Common Crawl, Colossal Clear Crawled Corpus C4 et The Pile. Le Common Crawl est un jeu de données constitué par une exploration à large échelle du Web. Il est notamment utilisé par OpenAI (Brown et al., 2020). Il est aussi utilisé comme *dataset* de base pour la constitution de *datasets* de meilleure qualité après l'application de règles de filtrage. C'est notamment le cas du Colossal Clear Crawled Corpus C4 (Dodge et al., 2021). Ce dernier s'appuie substantiellement sur les éditeurs de presse (New York Times, LA Times, Washington Post…) et les éditeurs scientifiques (PLOS One, Frontiers…), en plus de Google Patent pour l'accès aux connaissances scientifiques. The Pile est un jeu de données de haute qualité incluant 22 sous-*datasets* (Gao et al., 2020). Parmi les jeux de données, deux sont de nature scientifique : ArXiv (8,96 % du poids total) et PubMed Central (14,40 % du total). Le premier est un serveur de *preprints*, le second, un répertoire de documents issus de la recherche médicale. Aucun des deux ne propose une information soumise à un processus strict de *peer reviewing*. Les *datasets* intègrent classiquement des données issues de Wikipédia (Dodge et al., 2021). Or, il apparaît que Wikipédia est un bon relais pour l'information scientifique publiée, d'une part, dans des journaux en *open access*, d'autre part, dans des journaux à facteur d'impact élevé, éventuellement protégés par *paywall* (Teplitskiy et al., 2017).

Tous les contenus scientifiques n'ont en effet pas la même valeur (Cabanac, 2024). Premièrement, un contenu scientifique peut avoir fait ou non l'objet d'une révision par les pairs. Un contenu publié dans une conférence ou une revue à comité de lecture bénéficiera donc d'un niveau de validation supérieur à un article en *preprint*. Deuxièmement, à l'intérieur même des conférences ou des journaux scientifiques, une hiérarchie existe, que les articles soient ou non publiés en *open access*. Par exemple, l'indicateur SCImago Journal Rank (SJR) offre un classement des revues scientifiques contenues dans la base de données Scopus. Il est basé sur une mesure, inspirée du Pagerank de Google (Cardon, 2013), qui tient compte à la fois du nombre de citations reçues par une revue et du prestige des revues d'où proviennent les citations. Pour les producteurs d'IAG, il ressort, d'une part, que les contenus scientifiques les plus accessibles ne sont pas nécessairement les meilleurs,

d'autre part, qu'il existe un risque que les comportements protecteurs des éditeurs soient d'autant plus forts qu'une revue scientifique est réputée pour sa haute qualité.

En complément des *datasets* publics, les producteurs de LLM construisent aussi des *datasets* internes. Ceux-ci sont alimentés par, soit leurs propres robots, soit des accords signés avec les éditeurs (Gibney, 2024). Dès lors, les données d'entraînement sont composées, d'une part, de données publiques, utilisées après filtrage, collectées sur le Web par des tiers (p. ex. Common Crawl) ou par les producteurs eux-mêmes (p. ex. GPTbot), d'autre part, de données achetées auprès des éditeurs. Gibney (2024) mentionne ainsi Taylor & Francis[8] (accord avec Microsoft) et Wiley (partenaire inconnu), pour l'édition scientifique, ainsi que Financial Times (accord avec OpenAI), pour la presse spécialisée. Le phénomène concernerait cependant davantage d'éditeurs[9] tels que Elsevier, Springer Nature et De Gruyter Brill (Kwon, 2024). Le Generative AI Licensing Agreement Tracking tente ainsi d'inventorier les accords souscrits, souvent sans identification des bénéficiaires. Cette politique d'octroi de licences de gré à gré motive aussi les blocages mis en œuvre par les éditeurs (Kwon, 2024). De manière plus surprenante, des données volées, par exemple issues de Library Genesis (LibGen), seraient ponctuellement utilisées, par Meta (modèles LlaMa) et OpenAI (modèles GPT) notamment[10]. Le *dataset* Books3[11] serait ainsi incriminé. La composition précise des jeux de données d'entraînement demeure donc au final une inconnue, excepté pour des projets open-sources comme LUCIE[12], où la transparence est au cœur du projet.

Notre revue de littérature nous permet de formuler les hypothèses suivantes, qui vont être testées dans la suite de l'article : (H1) les robots des IA génératives sont davantage bloqués que les robots des moteurs de recherche ; (H2) le robot GPTbot est davantage bloqué que les robots d'autres IA génératives ; (H3) les éditeurs scientifiques commerciaux dominants bloquent davantage les robots d'IA génératives que les autres éditeurs ; (H4) les revues prédatrices bloquent moins les robots d'IA génératives que les revues non prédatrices ; et (H5) mieux une revue scientifique est classée et plus elle bloque les robots d'IA génératives.

3. Méthodologie

Deux jeux de données sont utilisés. Le premier est constitué de la liste de revues prédatrices publiée par Beall, et disponibles sur le site Beall's List. Le second est constitué des revues évaluées dans le Norwegian Register for Scientific Journals, Series and Publishers. Les revues y sont classées sur trois niveaux (level 0, level 1, level 2). Le niveau 1 intègre des revues scientifiques respectant les critères de

[8] Voir https://www.ccn.com/news/technology/microsoft-taylor-francis-secret-ai-publishing-deal-outrages-academics/ et https://www.informa.com/globalassets/documents/investor-relations/2024/informa-plc---market-update.pdf pour plus d'informations.

[9] L'exploitation des contenus scientifiques pourrait être facilitée par la structuration des articles selon le format *Journal Article Tag Suite* (JATS) Voir Kleidermacher et Zou (2025) pour un exemple d'exploitation du format JATS par une IA générative.

[10] Voir https://www.theatlantic.com/technology/archive/2025/03/libgen-meta-openai/682093/ (incluant les liens vers les dossiers juridiques) et https://authorsguild.org/news/you-just-found-out-your-book-was-used-to-train-ai-now-what/ pour plus d'informations.

[11] Voir par exemple https://aicopyright.substack.com/p/the-books-used-to-train-llms pour une analyse du contenu de ce jeu de données.

[12] Voir https://lucie.chat/ et https://huggingface.co/collections/OpenLLM-France/lucie-llm-67099ba7b992dee2c32b1f92 pour plus d'informations.

qualité académique élémentaires tandis que le niveau 2 rassemble les meilleurs canaux de publications. Le niveau 0 contient des revues non scientifiques, dédiées par exemple à la vulgarisation, mais aussi des revues prédatrices, telles que « Progress in Physics », également incluse dans la liste de Beall. L'appartenance éventuelle au Directory of Open Access Journals (DOAJ) y est indiquée. Parmi les classements publiés, celui de 2024 a été utilisé. Parmi ces listes, certains sites sont cependant injoignables, inaccessibles ou associés à des redirections (avec un envoi correct ou non du code HTTP correspondant). De plus, de nombreux doublons existent. En effet, plusieurs revues peuvent être publiées sur le même site (cas des grands éditeurs par exemple). Aussi un filtrage des URL (suppression des sites injoignables, calcul des redirections…) est réalisé à l'aide d'un script codé en Python. Au final ont été considérés 1153 sites de revues prédatrices, 4129 de niveau 0, 7276 de niveau 1 et 542 de niveau 2.

Les bases de données obtenues en sortie sont ensuite utilisées pour collecter les fichiers *robots.txt* puis, après analyse de ces derniers, obtenir un Top 10 des robots les plus fréquemment cités, identifier les pratiques de blocage (pas de fichier *robots.txt*, liste blanche, liste noire…) et calculer le taux de blocage par robot ainsi que le biais global. Le biais global est une valeur comprise entre -1 et 1 qui « *représente le pourcentage, en valeur absolue, de sites (de l'échantillon) qui favorisent (signe positif) ou défavorisent (signe négatif) le robot* » (Viseur & Delcoucq, 2024 ; Sun et al., 2007). Il est ainsi possible d'estimer la discrimination des robots d'IAG comparativement aux robots des moteurs de recherche. Le calcul du biais global utilise la procédure simplifiée définie par Viseur et Delcoucq (2024). Les résultats sont enregistrés dans un fichier journal. Ces fichiers peuvent ensuite être ingérés par ChatGPT pour la production de tableaux de synthèse spécifiques.

4. Résultats

H1 : Les robots des IA génératives sont davantage bloqués que les robots des moteurs de recherche.

Les robots les plus fréquemment cités par les revues non prédatrices sont, juste après le robot universel (« * »), GPTbot, CCbot, Google-Extended, Googlebot et ChatGPT-User. Les trois premiers robots sont les robots d'exploration utilisés pour la création et la mise à jour des jeux de données.

Tableau 1. Taux de blocage des robots d'exploration

Robot	Citations	Blocages	Taux de blocage (si robot cité)	Taux de blocage (tous les sites)
googlebot	689	9	1,31 %	0,08 %
bingbot	367	26	7,08 %	0,23 %
ccbot	763	666	87,29 %	5,96 %
gptbot	921	834	90,55 %	7,46 %
chatgpt-user	679	603	88,21 %	5,39 %
google-extended	720	657	91,25 %	5,88 %

Le blocage des robots d'exploration des deux moteurs de recherche dominants (Google et Bing) par les revues non prédatrices apparaît sensiblement moindre que celui des robots d'exploration des producteurs d'IA génératives (cf. Tableau 1). Même ChatGPT-User, le robot associé aux actions dans ChatGPT ou dans les « customs » ChatGPT fait l'objet d'un blocage fréquent. Surtout, dès lors que le robot est cité, c'est dans l'immense majorité des cas pour être finalement bloqué.

H2 : Le robot GPTbot est davantage bloqué que les robots d'autres IA génératives.

Le robot GPTbot fait l'objet d'un blocage par les revues non prédatrices dans 90,55 % (cf. Tableau 1) des cas où il est mentionné dans le fichier *robots.txt* (81,68 % des sites configurent un tel fichier). Au final, 7,46 % des sites web interdisent l'accès aux pages de contenu. Cette valeur est légèrement plus élevée (9,44 %) pour les sites des revues estampillées DOAJ.

Les autres robots d'exploration des IA génératives font l'objet d'un taux de blocage légèrement inférieur même si l'ordre de grandeur est équivalent. Le biais global pour GPTbot, soit -0,0743, est le plus élevé, et traduit une discrimination du robot comparativement à d'autres robots poursuivant ou non les mêmes objectifs de collecte.

H3 : Les éditeurs scientifiques internationaux bloquent davantage les robots d'IA génératives.

Les éditeurs internationaux comme Sciencedirect ou Springer ont un taux de blocage très sensiblement plus élevé que les revues prédatrices ou même que la moyenne des revues non prédatrices. Ce blocage accru conduit à un biais global (cf. Tableau 2) sensiblement plus élevé, en particulier pour le robot d'exploration GPTbot.

Tableau 2. Biais global (revues prédatrices vs Top 50)

Robot	Revues prédatrices	Top 50
googlebot	-0,0141	0,0323
bingbot	-0,0247	0,0323
ccbot	0	-0,1290
gptbot	-0,0018	-0,4194
chatgpt-user	0	-0,1613
google-extended	0	-0,2581

Les politiques de blocage des robots peuvent prendre des allures parfois radicales à l'image de Sciencedirect qui renvoie une erreur 403 (« *forbidden* ») lors de la lecture avec un script du fichier *robots.txt*. En pratique, le fichier existe (un hyperlien non fonctionnel déclencherait une erreur 404) mais son accès est activement bloqué après détection du robot. Ce dernier précise d'emblée la politique : « *# go away ? tell all others not in the list below to stay out!* ». Ce

dispositif pourrait s'expliquer par la volonté de freiner l'exploration des sites à large échelle et de limiter l'identification de ressources protégées par le droit d'auteur.

H4 : Les revues prédatrices bloquent moins les robots d'IA génératives que les revues non prédatrices.

Les revues prédatrices se distinguent par, d'une part, le plus faible pourcentage de sites disposant d'un fichier *robots.txt*, d'autre part, par le très faible biais global associé aux robots d'IA générative (cf. Tableau 3). Le biais global augmente sensiblement pour les revues de niveau 2.

Tableau 3. Biais global par type de revue

Robot	Revue prédatrice	Revue (niveau 0)	Revue (niveau 1)	Revue (niveau 2)
googlebot	-0,0141	-0,0027	0,0082	0,0112
bingbot	-0,0247	-0,0012	0,0082	0,0112
ccbot	0	-0,0194	-0,0551	-0,1453
gptbot	-0,0018	-0,0334	-0,0700	-0,2682
chatgpt-user	0	-0,0147	-0,0467	-0,1415
google-extended	0	-0,0167	-0,0529	-0,1750

H5 : Mieux une revue scientifique est classée et plus elle bloque les robots d'IA génératives.

Les revues non prédatrices ont une politique de régulation des robots d'exploration d'autant plus systématique que la revue est d'un niveau plus élevé. Cela se marque par l'utilisation plus systématique d'un fichier `robots.txt` (cf. Tableau 4).

Tableau 4. Utilisation du protocole d'exclusion des robots

Robot	Nombre de sites	Nombre de sites avec robots.txt
Revues prédatrices	1134	852 (75,1 %)
Revues de niveau 0	4070	3262 (80,1 %)
Revues de niveau 1	7169	5845 (81,5 %)
Revues de niveau 2	537	486 (90,5 %)

Tableau 5. Taux de blocage en fonction du niveau

Robot	Taux de blocage (ccbot)	Taux de blocage (google-extended)	Taux de blocage (gptbot)
Revues prédatrices	0 %	0 %	0,18 %
Revues de niveau 0	1,89 %	1,77 %	3,39 %
Revues de niveau 1	5,36 %	5,33 %	7,04 %
Revues de niveau 2	15,08 %	17,88 %	27,37 %
Top 50	19,35 %	32,26 %	48,39 %

Le taux de blocage augmente avec le niveau de la revue, légèrement jusqu'au niveau 1 puis plus brutalement pour les revues de niveau 2 (cf. Tableau 5). De plus, plus la revue est bien classée et plus le biais global présente une valeur négative élevée (cf. Tableau 3). Les revues de niveau 2 se distinguent particulièrement des revues de niveau 0 ou 1.

Toutes nos hypothèses (H1, H2, H3, H4 et H5) sont donc corroborées par les valeurs calculées des taux de blocage et des biais globaux.

5. Discussion

Hannigan et ses co-auteurs (2024) rappellent que « *les chatbots génératifs ne s'intéressent pas à la connaissance intelligente mais à la prédiction* ». Les grands modèles de langage (LLM, *Large Language Model*) sont en effet entraînés sur de vastes ensembles de données, souvent collectées à partir du Web, pour prédire des contenus. Ils sont capables de générer « *un charabia technique basé sur des motifs de mots dans les données d'entraînement, qui sont elles-mêmes une boîte noire* » (Hannigan et al., 2024). Deux choses méritent d'être soulignées à ce stade. D'une part, les grands modèles de langage ne disposent pas de la capacité à dégager un consensus scientifique par la compréhension profonde d'un corpus de documents. D'autre part, la qualité des contenus prédits dépend fortement de la qualité des données fournies en entraînement.

Or, si les robots d'exploration sont bloqués par les revues scientifiques bien classées et les grandes plateformes comme Springer ou ScienceDirect, le contenu provenant de ces sources de haute qualité risque d'être sous-représenté dans les jeux de données obtenues par *scraping*. En revanche, les revues prédatrices, qui ne bloquent pas ces robots, risquent d'y voir leur contenu surreprésenté. Ce déséquilibre entraîne un risque de diminution de la qualité de l'information scientifique que les *chatbots* (ou d'autres applications génératives) peuvent fournir. Les revues prédatrices publient souvent des articles sans processus rigoureux de validation par les pairs. Les LLM entraînés sur ces données sont davantage susceptibles de produire des informations erronées. Cette tendance est amplifiée par la moindre accessibilité des données issues des revues les mieux classées. Dès lors, en relayant des informations issues de sources peu fiables, les *chatbots* peuvent involontairement contribuer à la propagation de la mésinformation scientifique

(« *botshit* »). Cela est particulièrement préoccupant dans des domaines sensibles comme la santé, l'environnement ou la technologie, où des informations erronées peuvent avoir des conséquences graves. En outre, cela peut influencer négativement la perception de certains sujets scientifiques.

La composition des jeux de données d'entraînement impacte l'existence de biais parmi les réponses des intelligences artificielles (Chu et al., 2024 ; Hannigan et al., 2024 ; Ferrara, 2023 ; Navigli & Conia, 2023). Navigli et Conia (2023) introduisent ainsi le « *biais de sélection de données* », qu'ils définissent comme « *le biais causé par le choix des textes qui composent un corpus d'entraînement* », en complément des biais sociaux (sexisme, âgisme, racisme...). Ce biais se produit lorsque « *les textes sont identifiés, ou lorsque les données sont filtrées et nettoyées* ». En lien avec ce biais de sélection de données, notre recherche nous permet d'enrichir la typologie de Ferrara (2023) par l'ajout d'un biais de validation, que nous définissons comme la surreprésentation parmi le corpus d'entraînement de données faiblement validées sur un plan scientifique. De son côté, Ferrara (2023) identifie sept types de biais affectant les réponses de ChatGPT : les biais démographiques, les biais culturels, les biais linguistiques, les biais temporels, les biais de confirmation ainsi que les biais idéologiques et politiques. Les défauts des données d'entraînement ne sont pas irréversibles. Ils supposent cependant un travail fastidieux de rééquilibrage des *datasets*, c'est-à-dire de filtrage des données problématiques (Navigli & Conia, 2023). Cela peut d'ailleurs être vecteur de nouveaux biais (voir Dodge et al., 2021, par exemple, concernant l'introduction de biais démographiques lors du filtrage de contenus jugés grossiers).

Les politiques différenciées de blocage par les éditeurs scientifiques peuvent-elles engendrer d'autres biais que le biais de validation précédemment discuté ? Les biais temporels paraissent les plus évidents. Les LLM souffre en effet d'un temps de création élevé. D'une part, les jeux de données nécessitent du temps pour être constitués puis traités (Navigli & Cornia, 2023). Ces délais peuvent être accrus par l'existence d'étapes de traitement manuel, qui encouragent par ailleurs la réutilisation de jeux de données plus anciens. D'autre part, l'entraînement de l'IAG est lourd et prend donc lui-même du temps. Au 30 octobre 2024, les données d'entraînement du modèle GPT 4o n'allaient pas au-delà d'octobre 2023[13]. Les articles publiés au cours de l'année écoulée sont dès lors inconnus pour ChatGPT. Par ailleurs, les articles plus anciens ne sont pas nécessairement numérisés. Par exemple, certains *datasets* sont récents, comme arXiv qui remonte à 1991[14].

Les revues prédatrices modifient également sensiblement la provenance géographique des publications scientifiques. L'analyse de la localisation des serveurs hébergeant les revues prédatrices et non prédatrices permet ainsi de mettre en évidence des disparités entre ces deux types de revue. La localisation des sites web a été déterminée avec un script Python basé sur la solution GeoLite2 de MaxMind. Les 10 localisations les plus fréquentes pour les revues prédatrices ont été conservées puis comparées aux revues listées (niveaux 0, 1 et 2). Cette localisation met en évidence une surreprésentation des revues indiennes parmi les revues prédatrices, ce qui est également constaté par Xia et al. (2017). De plus, les revues non prédatrices ressortent comme globalement moins concentrées sur quelques pays, soit les États-Unis d'Amérique et l'Inde (plus de 50 % des revues

[13] Voir https://platform.openai.com/docs/models/gpt-4o.
[14] Voir https://en.wikipedia.org/wiki/ArXiv.

prédatrices). Dans les deux cas, le déséquilibre géographique est source de biais démographiques et de biais culturels, sans que l'impact soit facilement évaluable.

Les effets induits par le degré variable de validation des données scientifiques n'est sans doute homogène dans l'ensemble des disciplines scientifiques. Larivière et al. (2015) mettent ainsi en évidence les différences de dépendance aux éditeurs scientifiques commerciaux en fonction des disciplines. Les sciences sociales apparaissent par exemple beaucoup plus affectées que la physique dès lors que cette dernière bénéficie du support de puissantes sociétés savantes qui conservent davantage de contrôle sur la diffusion de la production scientifique. Par ailleurs, la diffusion des connaissances sur Wikipédia, qui peut servir de *proxy* pour l'accès à la connaissance scientifique derrière *paywall*, n'est pas homogène non plus pour tous les domaines (Teplitskiy et al., 2017). Il en résulte que les risques de mésinformation scientifique au sein des IAG varient probablement en fonction de la discipline.

Parmi les jeux de données à orientation scientifique, nous avons aussi vu que l'usage des *preprints* était répandu. Ce choix, notamment dictés par les facilités d'accès, est-il pénalisant du point de vue de la qualité des données collectées ? Dans le domaine de l'informatique, les pratiques de diffusion de recherches sous la forme de *preprints* sur arXiv a été étudiée par Lin et ses co-auteurs (2020). Leur recherche a nécessité un travail complexe de réconciliation des *preprints* et des versions publiées dans des actes de conférences ou des revues à comité de lecture. Leur recherche montre que près de 80 % des articles identifiés sont publiés dans un second temps. Les différences constatées concernent « *des révisions adéquates, des auteurs multiples, un résumé et une introduction détaillés, des références étendues et faisant autorité et un code source disponible* » (Lin et al., 2020). L'étude note cependant une tendance à la baisse du taux de publication des *preprints* (passé de 80 % à 75 % en quelques années). Ce haut taux de publication laisse cependant préjuger de la bonne qualité globale des recherches initiales. L'étude de Carneiro et ses co-auteurs (2020), appliquée à la littérature biomédicale (bioRxiv, PubMed), va dans le même sens. Les auteurs notent ainsi que la qualité des rapports est équivalente. Un léger avantage en faveur des versions publiées est cependant souligné. Même dans le contexte de la pandémie, propice à la mésinformation voire à la désinformation, les *preprints* sont apparus comme un allié précieux, par exemple pour la compréhension des mécanismes de transmission (Majumder & Mandl, 2020). Les *preprints* semblent donc ressortir comme un moyen efficace pour alimenter les jeux de données en informations scientifiques fiables et récentes.

Pour terminer, les problèmes de qualité des *datasets* peuvent également s'analyser, et être synthétisés, sous l'angle de la pertinence des représentations définie en management des systèmes d'information (Reix et al., 2011). Les données sont-elles, pour l'utilisateur, une représentation fiable de l'état de la connaissance à un instant donné ? Le premier problème est celui de l'accessibilité puisque certaines sources sont inaccessibles aussi bien pour les robots d'exploration collectant les données que pour les utilisateurs interagissant avec le *chatbot*. Surtout, ces *datasets* ne sont pas accessibles à des fins d'audit. Le second est celui de la fiabilité. Les données les plus facilement accessibles pour entraîner les intelligences artificielles génératives ne sont pas nécessairement les plus fiables comme le montrent les taux de blocage des éditeurs de presse ou des plateformes de l'édition scientifique. Cette question de la fiabilité est donc liée à celle de l'exactitude et de l'exhaustivité. La disponibilité accrue des contenus problématiques, couplée à la tendance naturelle aux hallucinations, conduit à des défauts d'exactitude. L'exhaustivité est impossible

dès lors qu'une partie de l'information échappe à l'utilisateur du fait de ces blocages. Le délai de constitution et de filtrage des jeux de données d'entraînement conduit à un défaut d'actualité. La mise à jour épisodique du modèle engendre par ailleurs un défaut de ponctualité. Dès lors, ces défauts de pertinence aboutissent à un ensemble d'erreurs et de biais dans les réponses.

Cette recherche souffre de cinq limitations. Premièrement, elle ne permet pas de nuancer les conclusions par discipline ou par famille de disciplines. Nous avons en effet vu, d'une part, que la dépendance aux éditeurs commerciaux dépendait du domaine de recherche, d'autre part, que des classements, notamment sectoriels, existaient. Le premier élément pourrait faciliter l'accès à des données d'entraînement, mais uniquement dans certaines disciplines, tandis que le second pourrait décourager la création de revues prédatrices dans des disciplines où des logiques de listes blanches prévalent dans l'évaluation des dossiers scientifiques (p. ex. classement français FNEGE[15] en science de gestion). Deuxièmement, le calcul de biais global a actuellement été réalisé sur l'ensemble des sites sans prendre en compte la concentration des revues sur quelques sites de grands éditeurs scientifiques (Springer, Sciencedirect...). Il en résulte une sous-estimation du biais global. Le calcul de ce dernier pour les 50 domaines les plus représentés en base de données donne cependant une indication de la borne supérieure du biais global pour les revues non prédatrices. Troisièmement, la recherche est limitée par l'utilisation de la liste de Beall. D'une part, la transparence de la méthodologie permettant de dresser cette liste a été critiquée (Richtig et al., 2018). D'autre part, cette liste est relativement ancienne puisque sa mise à jour a été stoppée en 2017 (Richtig et al., 2018). Aussi serait-il intéressant de recalculer la mesure de biais global sur une liste de revues prédatrices, faisant consensus et surtout plus récente. Nous pensons par exemple aux listes d'éditeurs et de journaux publiées sur le site predatoryjournals.org. Quatrièmement, la méthode retenue analyse une partie seulement des données accessibles aux producteurs pour entraîner leurs LLM. En effet, à côté des jeux de données publics, les développeurs construisent des jeux de données internes, notamment alimentés par des données achetées auprès des éditeurs (Gibney, 2024). L'ampleur de ces acquisitions semble actuellement limitée (Kwon, 2024). Cependant, elle est difficile à estimer en pratique dès lors que les contractants communiquent peu ou prou sur les accords. Cet accès privilégié aux données limite donc le biais de sélection pour certains modèles entraînés par des *bigtechs* (Google, Microsoft, OpenAI...). Le problème demeure cependant entier pour des modèles entraînés par des acteurs plus modestes ne disposant pas de cet accès privilégié aux données. Cette situation illustre par ailleurs le manque de transparence sur les données utilisées, dès lors la difficulté d'évaluer les risques inhérents à l'utilisation d'un modèle. Cinquièmement, seul le blocage passif par protocole d'exclusion est pris en compte par nos mesures. Or, nous avons vu que les éditeurs disposaient d'une vaste panoplie de dispositifs de blocage. Cependant, cette approche nous semble fiable dès lors que l'*opt-out* représente un premier niveau de régulation, adapté aux robots éthiques, préalable à l'utilisation de méthodes plus sophistiquées.

6. Conclusion

L'étude met en lumière l'impact des restrictions d'accès appliquées par les éditeurs scientifiques aux robots d'exploration des intelligences artificielles

[15] Voir https://fnege.org/classement-des-revues-scientifiques-en-sciences-de-gestion/.

génératives. D'une part, l'accès aux contenus intégraux est souvent bloqué par *paywall* ; d'autre part, l'accès aux sites des revues, donc des résumés des articles (si ces derniers sont derrière *paywalls*), est interdit par usage du protocole d'exclusion des robots. Nos résultats montrent le risque d'une prépondérance de contenus issus de revues prédatrices dans les données d'entraînement des IAG, créant ainsi un biais de validation. S'il n'est pas corrigé, par exemple via l'acquisition de contenus sous licence, ce biais expose les utilisateurs à une mésinformation scientifique, potentiellement amplifiée dans des domaines sensibles. La recherche souligne l'urgence de développer des stratégies de rééquilibrage des *datasets* en favorisant un accès contrôlé et éthique aux contenus validés par des pairs, afin d'améliorer la qualité et la fiabilité des réponses fournies par les modèles d'IAG. Notre recherche contribue ainsi à l'identification des biais dans les LLM ainsi qu'à leur mesure, leur compréhension et leur évitement (Chu et al., 2024 ; Ferrara, 2023 ; Navigli & Conia, 2023).

Cette recherche présente deux perspectives. Premièrement, la recherche actuelle ne particularise pas ses conclusions en fonction des disciplines scientifiques. Le taux de blocage des robots des producteurs d'IAG est-il homogène parmi l'ensemble de ces disciplines, ou bien certaines disciplines sont-elles davantage touchées que d'autres, et dès lors davantage exposées au risque de mésinformation scientifique ? Deuxièmement, le biais de validation a fait l'objet d'une estimation au niveau de la constitution des jeux de données brutes. L'analyse n'a pas été poussée jusqu'à des jeux de données filtrées. La contamination par des contenus de faible qualité, voire prédateurs, se vérifie-t-elle dans les jeux de données réellement utilisés, et dans quelles proportions ?

7. Références

Amin Azad, B., Starov, O., Laperdrix, P., & Nikiforakis, N. (2020). Web runner 2049: Evaluating third-party anti-bot services. In Detection of Intrusions and Malware, and Vulnerability Assessment: 17th International Conference, DIMVA 2020, Lisbon, Portugal, June 24–26, 2020, Proceedings 17 (pp. 135-159). Springer International Publishing. https://doi.org/10.1007/978-3-030-52683-2_7.

Banks, M. (2016). What Sci-Hub is and why it matters. American Libraries, 47(6), 46-49. https://www.jstor.org/stable/26380679.

Beall, J. (2010). "Predatory" open-access scholarly publishers. The Charleston Advisor, 11(4), 10-17.

Bontridder, N., & Poullet, Y. (2021). The role of artificial intelligence in disinformation. Data & Policy, 3, e32. https://doi.org/10.1017/dap.2021.20.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165. https://doi.org/10.48550/arXiv.2005.14165.

Cabanac, G. (2024). Fake Science: Misconduct Galore and Proposed Counterattack. Doctoral. IPBS, France. 2024. https://ut3-toulouseinp.hal.science/hal-04129541/.

Cardon, D. (2013). Dans l'esprit du PageRank: une enquête sur l'algorithme de Google. Réseaux, (1), 63-95. https://shs.cairn.info/revue-reseaux-2013-1-page-63.

Carneiro, C. F., Queiroz, V. G., Moulin, T. C., Carvalho, C. A., Haas, C. B., Rayêe, D., ... & Amaral, O. B. (2020). Comparing quality of reporting between preprints and peer-

reviewed articles in the biomedical literature. Research Integrity and Peer Review, 5, 1-19. https://doi.org/10.1186/s41073-020-00101-3.

Chawla, D. S. (2017). Publishers take academic networking site to court. Science, vol. 358, issue 6360, p. 161. https://www.science.org/doi/pdf/10.1126/science.358.6360.161.

Chu, Z., Wang, Z., & Zhang, W. (2024). Fairness in large language models: A taxonomic survey. ACM SIGKDD explorations newsletter, 26(1), 34-48. https://doi.org/10.1145/3682112.3682117.

Dinzinger, M., & Granitzer, M. (2024). A longitudinal study of content control mechanisms. In Companion Proceedings of the ACM on Web Conference 2024 (pp. 1382-1387). https://doi.org/10.1145/3589335.3651893.

Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., ... & Gardner, M. (2021). Documenting large webtext corpora: A case study on the colossal clean crawled corpus. arXiv preprint arXiv:2104.08758. https://doi.org/10.48550/arXiv.2104.08758.

Ferrara, E. (2023). Should chatgpt be biased? Challenges and risks of bias in large langage models. arXiv preprint arXiv:2304.03738. https://doi.org/10.5210/fm.v28i11.13346.

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. Minds and Machines, 30, 681-694. https://doi.org/10.1007/s11023-020-09548-1.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., ... & Leahy, C. (2020). The pile: An 800gb dataset of diverse text for language modeling. arXiv preprint arXiv:2101.00027. https://doi.org/10.48550/arXiv.2101.00027.

Gershenson, S., Polikoff, M. S., & Wang, R. (2020). When paywall goes AWOL: The demand for open-access education research. Educational Researcher, 49(4), 254-261. https://doi.org/10.3102/0013189X20909834.

Gibney, E. (2024). Has your paper been used to train an AI model? Almost certainly. Nature, 632(8026), 715-716. https://doi.org/10.1038/d41586-024-02599-9.

Hannigan, T. R., McCarthy, I. P., & Spicer, A. (2024). Beware of botshit: How to manage the epistemic risks of generative chatbots. Business Horizons. https://doi.org/10.1016/j.bushor.2024.03.001.

Kinney, R., Anastasiades, C., Authur, R., Beltagy, I., Bragg, J., Buraczynski, A., ... & Weld, D. S. (2023). The semantic scholar open data platform. arXiv preprint arXiv:2301.10140. https://doi.org/10.48550/arXiv.2301.10140.

Kleidermacher, H. C., & Zou, J. (2025). Science Across Languages: Assessing LLM Multilingual Translation of Scientific Papers. arXiv preprint arXiv:2502.17882. https://doi.org/10.48550/arXiv.2502.17882.

Kwon, D. (2024). Publishers are selling papers to train AIs-and making millions of dollars. Nature, 636(8043), 529-530. https://doi.org/10.1038/d41586-024-04018-5.

Larivière, V., Haustein, S., & Mongeon, P. (2015). The oligopoly of academic publishers in the digital era. PloS one, 10(6), e0127502. https://doi.org/10.1371/journal.pone.0127502.

Lin, J., Yu, Y., Zhou, Y., Zhou, Z., & Shi, X. (2020). How many preprints have actually been printed and why: a case study of computer science preprints on arXiv. Scientometrics, 124(1), 555-574. https://doi.org/10.1007/s11192-020-03430-8.

Lo, K., Wang, L. L., Neumann, M., Kinney, R., & Weld, D. S. (2020). S2ORC: The semantic scholar open research corpus. arXiv preprint arXiv:1911.02782. https://doi.org/10.48550/arXiv.1911.02782.

Maleki, N., Padmanabhan, B., & Dutta, K. (2024). AI hallucinations: a misnomer worth clarifying. In 2024 IEEE Conference on Artificial Intelligence (CAI) (pp. 133-138). IEEE. https://doi.org/10.1109/CAI59869.2024.00033.

Majumder, M. S., & Mandl, K. D. (2020). Early in the epidemic: impact of preprints on global discourse about COVID-19 transmissibility. The lancet global health, 8(5), e627-e630. https://doi.org/10.1016/S2214-109X%2820%2930113-3.

Navigli, R., Conia, S., & Ross, B. (2023). Biases in large language models: origins, inventory, and discussion. ACM Journal of Data and Information Quality, 15(2), 1-21. https://doi.org/10.1145/3597307.

Reix, R., Fallery, B., Kalika, M., & Rowe, F. (2011). Systèmes d'information et management des organisations. Vuibert. ISBN : 9782711743810.

Richtig, G., Berger, M., Lange-Asschenfeldt, B., Aberer, W., & Richtig, E. (2018). Problems and challenges of predatory journals. Journal of the European Academy of Dermatology and Venereology, 32(9), 1441-1449. https://doi.org/10.1111/jdv.15039.

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-J. P., & Wang, K. (2015). An overview of Microsoft Academic Service (MAS) and applications. Proceedings of the 24th International Conference on World Wide Web (WWW '15 Companion), 243-246. ACM. https://doi.org/10.1145/2740908.2742839.

Sun, Y., Zhuang, Z., Councill, I. G., & Giles, C. L. (2007). Determining bias to search engines from robots. txt. In IEEE/WIC/ACM International Conference on Web Intelligence (WI'07) (pp. 149-155). IEEE. https://doi.org/10.1109/WI.2007.98.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). ArnetMiner: Extraction and mining of academic social networks. Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'2008), 990-998. https://doi.org/10.1145/1401890.1402008.

Teplitskiy, M., Lu, G., & Duede, E. (2017). Amplifying the impact of open access: Wikipedia and the diffusion of science. Journal of the Association for Information Science and Technology, 68(9), 2116-2127. https://doi.org/10.1002/asi.23687.

Viseur, R., & Delcoucq, L. (2024). Exploration des pratiques de régulation des IA génératives par le protocole d'exclusion des robots. INFORSID, 28-31 mai 2024, Nancy (France). http://inforsid.fr/actes/2024/inforsid24-89-104.pdf.

Viseur, R. (2024). Analyse de l'impact des IA génératives sur la presse en ligne : anatomie d'un newsbot basé sur GPT. Actes des conférences AIM. Montpellier (France). https://aim.asso.fr/fr/publications/actes-conferences/id-1809-aim2024_557858.

Wang, K., Shen, Z., Huang, C., Wu, C. H., Dong, Y., & Kanakia, A. (2020). Microsoft academic graph: When experts are not enough. Quantitative Science Studies, 1(1), 396-413. https://doi.org/10.1162/qss_a_00021.

Xia, J., Li, Y., & Situ, P. (2017). An overview of predatory journal publishing in Asia. Journal of East Asian Libraries, 2017(165), 4. https://scholarsarchive.byu.edu/jeal/vol2017/iss165/4.

Xia, J., Harmon, J. L., Connolly, K. G., Donnelly, R. M., Anderson, M. R., & Howard, H. A. (2015). Who publishes in "predatory" journals?. Journal of the Association for Information Science and Technology, 66(7), 1406-1417. https://doi.org/10.1002/asi.23265.

Ye, H., Liu, T., Zhang, A., Hua, W., & Jia, W. (2023). Cognitive mirage: A review of hallucinations in large language models. arXiv preprint arXiv:2309.06794. https://doi.org/10.48550/arXiv.2309.06794.

Résolution d'entités pour les flux de données à l'aide de la technique d'embedding

Zhongwei MA [1], **Philippe ROOSE** [1], **Jiefu SONG** [2]

1. *Universite de Pau et des Pays de l'Adour, E2S UPPA, LIUPPA Anglet, France*
   zhongwei.ma@univ-pau.fr,philippe.roose@univ-pau.fr
2. *Institut de Recherche en Informatique de Toulouse - Université Toulouse Capitole 31000, Toulouse, France*
   jiefu.song@ut-capitole.fr

RÉSUMÉ. *La plupart des systèmes de traitement de flux de données recueillent des données provenant de différentes sources en temps réel et les consomment immédiatement. Cependant, de nombreuses analyses décisionnelles nécessitent des données en temps réel et des données historiques (par exemple, un dataset local ou des enregistrements antérieurs) en même temps pour comprendre la situation actuelle et avoir une vue globale. La résolution d'entités permet de déterminer si deux enregistrements différents font référence à la même entité en l'absence d'un identifiant unifié dans le cas multi-sources. Il est donc essentiel d'intégrer les données en temps réel aux données stockées en appliquant la résolution d'entités et de garantir l'accessibilité et la facilité d'utilisation des données multi-sources. Les méthodes existantes pour la résolution d'entités sont souvent incapables de prendre en charge ces dernières de manière continue tout en fournissant des méthodes générales et efficaces pour l'analyse de données complexes telles que le texte. Étant donné la bonne capacité de l'embedding dans la capture des informations sémantiques et syntaxiques, nous visons à appliquer cette technique au traitement de flux de données dans le but d'intégrer les données entrants en temps réel aux données historiques. En outre, la plupart des approches privilégient aujourd'hui la précision et la scalabilité tout en ignorant l'augmentation de la consommation d'énergie que nécessite l'amélioration de cette précision. Nous proposons donc ici une approche d'embedding de graphe dynamique adaptée au traitement des données en temps réel pour effectuer la résolution d'entités dans des tables relationnelles tout en évaluant la consommation d'énergie de ce traitement.*

ABSTRACT.

MOTS-CLÉS : *flux de données, résolution d'entité, embedding*

KEYWORDS: *entity resolution, data stream, embedding*

1. Introduction

La résolution d'entités (Entity Resolution, ci-après dénommé ER), également appelée couplage d'enregistrements (record linkage) ou l'alignement de données (data matching), vise à trouver différentes descriptions se rapportant à la même entité dans le monde réel, au sein de sources de données ou entre celles-ci, lorsqu'il n'existe pas d'identifiant unique de l'entité, et ceci afin de garantir la qualité et la cohérence des datasets intégrés. (Christophides *et al*., 2020). En supposant que nous disposions d'enregistrements de ventes en temps réel sur différentes plateformes de distributeurs, si ces enregistrements sont intégrés aux enregistrements historiques de notre base de données, nous pouvons obtenir des informations plus approfondies sur les produits populaires et les tendances d'achat des consommateurs. Il s'agit d'une étape cruciale lorsque nous procédons à l'intégration de données et que nous voulons offrir une vue uniforme de sources de données autonomes et hétérogènes, ce qui facilite le traitement ultérieur (Maharana *et al*., 2022).

Dans un contexte de big data, les données, caractérisées par leur nature continue et temps réel, sont de plus en plus utilisées dans le monde numérique (Bahri *et al*., 2021). Ces données sont générées à partir d'un large éventail de sources, notamment les plateformes de réseaux sociaux, les objets de l'IoT, etc. Les exigences en matière d'analyse des flux de données sont de deux aspects : le traitement en temps réel pour soutenir la prise de décision immédiate avec une faible latence, et le stockage des données pour des requêtes ultérieures ou des analyses plus complexes et plus coûteuses en temps. Par conséquent, les systèmes modernes de traitement des flux de données doivent intégrer à la fois le traitement en flux et le traitement par lots (Isah *et al*., 2019) dans un cadre unifié, comme l'architecture Lambda introduite à l'origine par (Marz, Warren, 2015). Pour améliorer l'efficacité de ces systèmes, il est essentiel d'intégrer les données de flux aux données historiques par ER.

Le premier défi tient à l'incomplétude des datasets. En raison de la nature illimitée des flux, un cadre dynamique est nécessaire pour exécuter l'ER de façon incrémentale. L'absence d'un dataset complet complique la tâche, car il devient difficile d'identifier la meilleure correspondance dans un volume de données fini.

Historiquement, l'ER a été définie comme une tâche hors ligne réalisée lors de l'intégration des données afin d'améliorer la qualité des bases de données (Christophides *et al*., 2020). De nombreuses approches ont ainsi été développées pour traiter ce problème en mode batch à l'aide de différentes méthodes. Cependant, ces approches orientées batch, conçues pour des bases de données statiques, ne sont pas adaptées aux flux de données en continu. Jusqu'à présent, deux principales approches ont été adaptées aux données incrémentales. La première repose sur des méthodes basées sur des règles (Gazzarri, Herschel, 2023) Toutefois, ces méthodes s'appuient souvent sur des règles complexes et sont limitées à des domaines spécifiques, ce qui restreint leur généralisabilité. La seconde approche repose sur l'apprentissage automatique, en particulier les techniques de clustering (Do Nascimento *et al*., 2018 ; Saeedi *et al*., 2020).

Néanmoins, cette approche rencontre des difficultés à traiter efficacement les données textuelles non structurées, ce qui limite son efficacité dans des environnements réels.

Le deuxième défi porte sur la consommation. La grande quantité de données et leurs traitements entraînent une consommation significative de ressources.

Pour surmonter les limitations de l'analyse textuelle et la complexité des algorithmes spécifiques à un domaine, nous nous tournons vers les techniques d'embedding, qui projettent les données textuelles dans un espace vectoriel de faible dimension afin de capturer les relations et l'information sémantique des mots (Wang *et al*., 2020). Jusqu'à présent, cette technique est principalement utilisée pour le traitement en batch (Li *et al*., 2020). Récemment, d'importants travaux ont été menés sur les embeddings incrémentaux (Barros *et al*., 2021), jetant ainsi les bases de l'ER dans un contexte de flux de données sans s'appuyer sur le traitement en batch.

Dans ce travail, nous introduisons un cadre pour l'intégration de données en flux continu basé sur l'embedding dynamique de graphes. Ce cadre englobe l'ensemble du processus, depuis l'entraînement préalable des modèles d'embedding sur des datasets locaux, jusqu'à l'intégration des données en flux, en impliquant progressivement les nouvelles données et en déterminant si elles appartiennent à une entité existante dans l'dataset d'origine. La construction des embeddings permet de capturer efficacement les informations sémantiques et syntaxiques (Wang *et al*., 2020). De plus, afin de relever les défis énergétiques, nous proposons une approche originale permettant de mesurer la consommation énergétique lors de l'exécution du code d'embedding pour l'intégration des données. Nous proposons les contributions suivantes :

1) Modèle d'embedding dynamique de graphes. Basé sur des modèles d'embedding capturant les relations et les informations sémantiques des données, nous mettons à jour le graphe et ses embeddings de manière incrémentale. Plus précisément, à l'arrivée d'un nouvel enregistrement, nous ajoutons un nœud correspondant au graphe, établissons des arêtes entre ce nœud et la structure existante, et générons des parcours aléatoires évolutifs à partir du nouveau nœud afin de quantifier la similarité. Cette approche évite l'entraînement répétitif, réduisant ainsi les coûts de calcul tout en permettant un ER efficace dans un contexte de flux de données.

2) Évaluation du modèle en termes de performance et d'efficacité. Nous évaluons la méthode proposée en analysant sa consommation énergétique ainsi que ses performances. Cette analyse permet une compréhension quantitative des ressources consommées dans les tâches d'ER basées sur l'embedding.

L'article est structuré comme suit. La section 2 présente les méthodes courantes d'ER et en analyse les caractéristiques. La section 3 introduit la conception de notre approche. Dans la section 4, nous décrivons en détail l'implémentation du protocole et analysons les résultats obtenus. Enfin, nous concluons dans la dernière section en évoquant les perspectives de travaux futurs.

2. État de l'art

Nous passons en revue la littérature liée à cet article selon deux axes. Dans **la Section 2.1**, nous résumons les méthodes adaptées aux flux de données afin de présenter l'état actuel de l'ER incrémental. **La Section 2.2** se concentre sur les techniques d'embedding et illustre comment elles peuvent être appliquées pour réaliser l'ER.

2.1. Résolution d'entités pour les flux

De nombreuses études mentionnées dans la littérature (Binette, Steorts, 2022) ont exploré diverses approches pour l'ER afin de relever le défi des données incrémentales. Nous les catégorisons en trois grands groupes : les approches basées sur les requêtes, les approches basées sur les règles et les approches basées sur l'apprentissage.

Les approches basées sur les requêtes (Simonini *et al*., 2022) stockent toutes les données entrantes et effectuent l'ER au moment de l'exécution de la requête. Cependant, ces méthodes à la demande peuvent nécessiter plusieurs minutes de traitement pour fournir un résultat. Par ailleurs, le stockage de doublons dans les enregistrements augmente également la consommation des ressources, entraînant une inefficacité.

Les approches basées sur les règles (Gazzarri, Herschel, 2021) sont parmi les plus couramment utilisées pour l'ER. Elles reposent sur des règles déterministes et interprétables permettant de comparer les attributs des enregistrements. Toutefois, la mise en œuvre de ces approches peut être très complexe et nécessite des algorithmes sophistiqués adaptés aux flux de données.

Les approches basées sur l'apprentissage sont devenues de plus en plus courantes grâce aux avancées des techniques d'apprentissage automatique *(machine learning)*. La plupart de ces méthodes fonctionnent en mode batch (Papadakis *et al*., 2023) et visent à améliorer la précision et l'efficacité grâce à des techniques telles que la correspondance de graphes et les réseaux neuronaux. Cependant, l'entraînement des modèles d'apprentissage automatique présente des défis importants, en particulier en présence de données avec bruits ou de ressources annotées limitées, et ce d'autant plus dans un contexte de flux de données où le dataset est incomplet. Les méthodes incrémentales adoptent souvent des techniques non supervisées comme le clustering (Saeedi *et al*., 2020) Plutôt que de lier individuellement les enregistrements, l'objectif du clustering est de regrouper les enregistrements correspondant aux mêmes entités, souvent latentes. Néanmoins, les performances des méthodes de clustering restent limitées lorsqu'il s'agit de traiter des données de grande dimension, telles que le texte, qui joue un rôle crucial dans l'ER. De plus, les données avec bruits, comme les fautes d'orthographe ou les valeurs manquantes, peuvent générer du bruit et des faux positifs qui diminuent la précision, en particulier dans le cas des flux de données (Christophides *et al*., 2020).

2.2. Embeddings et résolution d'entités

L'embedding de mots est une technique clé en apprentissage automatique, visant à projeter des données de haute dimension dans un espace vectoriel où les mots sont représentés par des vecteurs de longueur fixe. Cette représentation permet de capturer les relations sémantiques entre mots similaires (Almeida, Xexéo, 2023).

Les approches basées sur la prédiction, par exemple, word2vec (Mikolov *et al*., 2013), entraînent des réseaux neuronaux récurrents en exploitant les données locales (par exemple, le contexte d'un mot) afin soit de prédire un mot à partir de son contexte, soit d'inférer le contexte à partir d'un mot donné. Ce processus garantit que les mots ayant des significations sémantiques similaires sont associés à des représentations vectorielles similaires. D'un autre côté, les méthodes basées sur le comptage (par exemple, Glove (Pennington *et al*., 2014)) construisent une matrice de cooccurrence de mots à partir d'un corpus, capturant des informations statistiques globales telles que le nombre total d'occurrences des mots et leurs fréquences. Voici quelques applications clés des techniques d'embedding dans ce domaine.

DeepER (Ebraheem *et al*., 2018) est l'un des premiers à utiliser des embeddings de mots (comme GloVe). Il applique des réseaux LSTM (Long Short-Term Memory) pour apprendre les relations entre les attributs des tuples en étiquetant les données et en les convertissant en une représentation vectorielle de dimensions fixes. Ensuite, des caractéristiques de similarité sont calculées et introduites dans un classificateur binaire afin de déterminer les correspondances entre entités. DeepMatch (Mudgal *et al*., 2018) adopte une approche similaire mais offre davantage de choix pour l'embedding des attributs et la représentation des similarités. Ils démontrent des performances compétitives sur des datasets contenant une certaine proportion de données avec bruits.

Ditto (Li *et al*., 2020) utilise des modèles pré-entraînés basés sur des "transformers" pour extraire des caractéristiques et ajuste le modèle en tant que classificateur binaire pour l'appariement d'entités. Il gère mieux les données avec bruits grâce à un processus de sérialisation structuré et à des techniques d'augmentation de données. (Brunner, Stockinger, 2020) teste également des modèles basés sur les transformers et explore les mécanismes d'attention pour la tâche d'ER. Cependant, l'utilisation de grands modèles de langage engendre des coûts énergétiques énormes en raison d'une exploitation de ressources significative.

Plus récemment, plusieurs frameworks ont proposé de représenter les données ou les paires d'enregistrements sur la base des embeddings, tels que multiEM (Zeng *et al*., 2024), Unicorn (Tu *et al*., 2023), et FlexER (Genossar *et al*., 2023). Ces frameworks emploient des techniques avancées, notamment les réseaux de neurones graphiques, afin d'identifier les similarités entre enregistrements et d'améliorer les capacités de généralisation. Néanmoins, ils encodent les données sous forme de séquences linéaires fixes basées sur des règles prédéfinies, ce qui limite leur capacité à interpréter les tables relationnelles, telles que les relations entre colonnes. Afin d'obtenir une interprétation plus complète, Embdi (Cappuzzo *et al*., 2020) introduit des représentations graphiques dans les tâches d'ER pour enrichir l'information capturée par les embeddings.

Malgré ces avancées, toutes les méthodes mentionnées fonctionnent en mode batch, ce qui limite leur application aux tâches nécessitant un traitement en temps réel. Pour traiter des flux de données en continu, il est nécessaire de disposer d'une méthode permettant d'intégrer progressivement les nouvelles données sans devoir relancer manuellement le programme à chaque fois.

3. Proposition

3.1. Problématique

Pour simuler l'intégration des données en flux dans un dataset local, nous supposons l'existence de deux sources de données : l'une contient l'ensemble des enregistrements issus d'une source statique, où chaque enregistrement correspond à une entité distincte, et l'autre est constituée des données entrantes en streaming. L'objectif de l'ER incrémental est d'associer un enregistrement entrant à l'entité la plus similaire dans un ensemble de références en évolution. Cet ensemble de références comprend à la fois le jeu de données statiques et les enregistrements de flux précédemment observés, et il se met à jour dynamiquement au fur et à mesure que de nouvelles données arrivent. Pour préciser davantage notre cadre, nous introduisons les principaux symboles dans la table 1 et les algorithmes utilisés dans les approches incrémentales.

TABLEAU 1. *Définitions des symboles*

Symbole	Explication
D	Un dataset statique, qui peut être vide
$\mathcal{D}$	Un dataset dynamique
ΔD	L'incrément d'un dataset incrémental pendant une période
R_i	Un enregistrement arbitraire
$\mathcal{C}_t$	L'ensemble des données traitées avant le temps t, servant d'entrée pour le prochain cycle de traitement
M_{t,R_i}	La meilleure correspondance pour l'enregistrement R_i au temps t
$\mathcal{M}_t$	La liste des correspondances pour tous les enregistrements au temps t
$sim(\cdot,\cdot)$	Une fonction de similarité pour comparer les enregistrements

L'ER dans les flux de données présente deux principaux défis. Le premier est lié à l'incomplétude de datasets dynamiques. À un instant t, nous n'avons qu'une vue partielle de $\mathcal{D}$, ce qui empêche d'attendre que l'ensemble des données soit disponible avant d'exécuter le processus. Plus précisément, nous pouvons représenter le dataset $\mathcal{D}$ à un instant donné comme suit :

$$\mathcal{D}_n = \bigsqcup_{i=1}^{n} \Delta D_i$$

où $\mathcal{D}_n$ représente le dataset formé en combinant les incréments $\Delta D_1, \ldots, \Delta D_n$, chaque incrément de données ΔD_i correspondant à de nouvelles données arrivant dans un intervalle de temps spécifique $[t_{i-1}, t_i]$. Nous supposons que ces intervalles sont séquentiels et non chevauchants, c'est-à-dire que chaque nouvel intervalle temporel suit *(meet)* immédiatement le précédent : $[t_{i-1}, t_i]$ meet $[t_i, t_{i+1}]$.

Ainsi, le dataset $\mathcal{D}_n$ évolue au fil du temps par l'ajout séquentiel de ces incréments.

Plutôt que de comparer l'ensemble des données en mode batch, ce qui est impossible étant donné que le flux de données est généré de manière continue sans fin, le processus doit fonctionner de manière incrémentale. Une approche raisonnable consiste à exécuter le processus d'ER à chaque arrivée d'un nouvel incrément ΔD_n, afin d'éviter une accumulation excessive de données. Le moment de ces mises à jour peut être déterminé par divers facteurs, tels qu'un nombre prédéfini de nouveaux enregistrements ou des intervalles de temps fixes.

L'idée principale de l'ER incrémental est de réutiliser les résultats des traitements précédents pour éviter les calculs redondants. Dans ce cadre, pour des données incrémentales ΔD_n à l'instant t, les données précédemment traitées sont stockées dans un ensemble de candidats, représenté par $\mathcal{C}_t = D \cup \mathcal{D}_{n-1}$. Ces données ont été transformées de leur état initial vers une forme prête pour la comparaison. À chaque mise à jour incrémentale, l'ensemble de candidats est continuellement enrichi avec les nouvelles données, tandis que les données précédemment traitées restent inchangées et sont uniquement utilisées pour la fonction de similarité. La Figure 1 illustre l'évolution des données impliquées dans ce processus incrémental au fil du temps. Ainsi, nous pouvons formuler le premier problème comme suit :

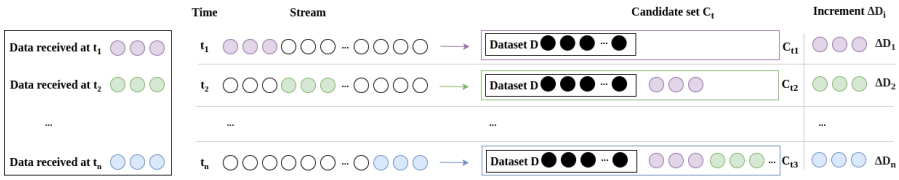


FIGURE 1. *Données impliquées dans le processus incrémental au fil du temps*

Définition du problème 1. À tout instant t, mettre à jour de manière incrémentale l'ensemble de candidats $\mathcal{C}_t$ et comparer les paires de données à l'aide de la fonction de similarité $sim(\Delta D_n, \mathcal{C}_t)$ afin de déterminer M_{t,R_i} pour chaque élément $R_i \in \Delta D_n$.

Le second problème est que la mise en correspondance n'est pas un processus statique dans le temps. Étant donné que les enregistrements candidats évoluent continuellement, les correspondances précédemment établies M_{t',R_i} (où $t' < t$) peuvent devoir être mises à jour lors de l'arrivée de nouveaux enregistrements. Cela nécessite un maintien continu de la cohérence des résultats de correspondance. Nous définissons l'ensemble des correspondances à l'instant t comme suit :

$$\mathcal{M}_t = \{(R_i, M_{t,R_i}) \mid i \leq t\}$$

Avec la croissance de $\mathcal{M}_t$, le stockage de toutes les correspondances historiques devient rapidement impraticable. Pour garantir l'efficacité, il est nécessaire de définir une stratégie permettant de déterminer quelles correspondances doivent être conservées ou supprimées. Nous résumons ce problème ainsi :

Définition du problème 2. Gérer et maintenir les résultats de correspondance dynamiques au fil du temps $\mathcal{M}_t$, de sorte qu'à tout instant t, la correspondance correcte apparaisse dans la liste dès qu'elle est identifiée, et que sa valeur de similarité reste relativement élevée par rapport à l'ensemble des enregistrements potentiellement similaires.

Nos pipelines répondent aux problèmes 1 et 2 via un processus d'ER incrémental qui met à jour efficacement les correspondances tout en réduisant l'espace de stockage. Toutes les informations de correspondance pertinentes sont stockées dans un index unique, assurant un accès uniforme aux requêtes. Les détails sont présentés dans la section suivante.

3.2. Proposition

Notre cadre d'ER incrémental [1] comprend la préparation des données, la construction d'embeddings incrémentaux et la génération des correspondances. Dans ce travail, nous réutilisons l'approche de construction d'un modèle d'embedding présentée dans (Cappuzzo *et al*., 2020), qui est adaptée à l'extension aux flux de données et représente des informations sémantiques riches dans un modèle linguistique léger. Cependant, au lieu de construire un modèle d'embedding basé sur deux datasets complets, nous introduisons un embedding incrémental qui met à jour le modèle au fil du temps afin d'améliorer la scalabilité des applications de traitement en flux. Pour illustrer comment ces trois parties s'intègrent et comment les données circulent entre elles, nous résumons ces étapes dans la section suivante.

3.2.1. Préparation des données

La préparation des données traite les incréments bruts ΔD_i provenant des sources de données qui génèrent des données en temps réel. L'extraction de métadonnées est effectuée à cette étape, permettant d'obtenir des descriptions concises et représentatives des entités du monde réel. Une table relationnelle est ensuite construite à partir des métadonnées extraites, fournissant une base flexible pour diverses tâches dans différents domaines. Dans ce travail, elle est spécifiquement utilisée pour l'ER. Cette étape de préparation des données transforme des données non structurées en un format structuré, réduisant efficacement le volume de données à comparer et standardisant les informations pour les traitements en aval.

3.2.2. Construction incrémentale des embeddings

À partir de la table relationnelle contenant les incréments de données, nous pouvons construire un modèle d'embedding incrémental. Le modèle utilisé dans ce processus est pré-entraîné sur un dataset local D, mais l'étape de pré-entraînement peut être omise si nous intégrons des flux de données à partir de zéro. Le pipeline d'entraînement est structuré comme suit : chaque enregistrement dans la table relationnelle se

1. https://github.com/Mzhongwei/er_embedding_streaming

Algorithme 1 : Algorithme d'Embedding Incrémental

Input : incrément de données ΔD_i, modèle d'embedding $\mathcal{E}_{i-1}$, graphe $\mathcal{G}_{i-1}$
Output : modèle d'embedding mis à jour $\mathcal{E}_i$, graphe mis à jour $\mathcal{G}_i$

1: Initialiser la liste des nœuds évolutifs $\mathcal{L}$
2: Initialiser le graphe $\mathcal{G}_i \leftarrow \mathcal{G}_{i-1}$
3: **foreach** $R \in \Delta D_i$ **do**
4: $\quad \mathcal{G}_i, \Delta N_1 \leftarrow$ ajouterNœudAGraphe$(\mathcal{G}_i, R)$
5: $\quad \mathcal{G}_i, \Delta N_2 \leftarrow$ ajouterLienAGraphe$(\mathcal{G}_i, R)$
6: $\quad \mathcal{L} \leftarrow$ ajouterNœudDynamiqueÀListe$(\mathcal{G}_i, \Delta N_1, \Delta N_2)$
7: Générer de nouvelles random walks pour $\mathcal{L}$: $\mathcal{W} =$ GenererRandomWalk$(\mathcal{G}, \mathcal{L})$;
8: Entraîner le modèle avec les random walks $\mathcal{W}$:
$\mathcal{E}_i =$ MettreAJourEmbedding$(\mathcal{E}_{i-1}, \mathcal{W})$
9: **return** $\mathcal{E}_i$, $\mathcal{G}_i$

voit attribuer un identifiant unique. Chaque valeur, y compris les numéros d'ID et les noms de colonnes, est traitée comme un nœud dans un graphe hétérogène non orienté. Un random walk est ensuite effectué sur ce graphe afin de générer des représentations contextuelles de ces valeurs et mots sous forme de phrases. Ces phrases sont ensuite projetées dans un espace vectoriel de dimension fixe, formant ainsi la représentation en embedding des valeurs (Cappuzzo *et al.*, 2020).

Le pipeline incrémental suit la même structure mais intègre dynamiquement les nouvelles données. L'ensemble du processus est présenté dans l'algorithme 1. Lorsqu'un incrément de données ΔD_i arrive au temps t_i, nous utilisons le graphe $\mathcal{G}_{i-1}$ et le modèle d'embedding $\mathcal{E}_{i-1}$ à l'instant précédent t_{i-1} comme variables initiales. Le graphe est mis à jour en ajoutant de nouveaux nœuds contenant les valeurs extraites de chaque enregistrement R de l'incrément ΔD_i, si ces valeurs ne sont pas déjà présentes. Ensuite, de nouvelles arêtes sont insérées dans le graphe pour représenter les relations entre différents enregistrements et différentes valeurs au sein d'un même enregistrement (cf. lignes 3-5). À cette étape, tous les nœuds ayant évolué, y compris les nouveaux nœuds ΔN_1 ainsi que ceux dont les arêtes ont changé ΔN_2, sont enregistrés dans une liste de nœuds évolutifs $\mathcal{L}$ (cf. ligne 6). Un nouveau random walk $\mathcal{W}$ est ensuite effectué à partir de ces nœuds pour générer de nouvelles phrases (cf. ligne 7). Ces phrases sont utilisées pour affiner le modèle d'embedding par apprentissage incrémental, garantissant que les représentations restent à jour sans nécessiter un réentraînement complet depuis le début (cf. ligne 8). Le graphe mis à jour $\mathcal{G}_i$ et le modèle d'embedding $\mathcal{L}_i$ sont stockés pour l'itération suivante.

3.2.3. Construction de la liste de correspondances

La construction de la liste de correspondances prend en entrée les identifiants des enregistrements incrémentaux pour représenter ces enregistrements et le modèle d'embedding renouvelle à l'étape précédente. Comme illustré dans l'algorithme 2, pour chaque enregistrement représenté par un identifiant, nous pouvons calculer les degrés de similarité entre les vecteurs des numéros d'ID afin d'identifier les enregistrements

les plus similaires selon les représentations vectorielles du modèle d'embedding (cf. ligne 2). Ensuite, nous sélectionnons les k correspondances les plus similaires (cf. ligne 3). Ces relations, ainsi que leurs scores de similarité, sont stockées symétriquement dans une liste, dont seules les n meilleures correspondances sont conservées par enregistrement (cf. lignes 4-6).

Algorithme 2 : Construction de la Liste de Correspondances

Input : Identifiants des enregistrements incrémentaux $\mathcal{ID} = \{id_1, id_2, \ldots, id_n\}$, modèle d'embedding $\mathcal{E}_i$, ensemble de candidats $\mathcal{C}$, nombre de correspondances les plus similaires k après calcul et n dans la liste de correspondances, liste de correspondances précédente $\mathcal{M}_{i-1}$

Output : Liste de correspondances mise à jour $\mathcal{M}_i$

1: **foreach** $id \in \mathcal{ID}$ **do**
2: Exécuter la fonction de similarité : Sim($\mathcal{E}_i$, id)
3: Sélectionner les k identifiants les plus similaires avec leur score de similarité s : $\mathcal{N}_{id} = \{(j_1, s_1), (j_2, s_2), \ldots, (j_k, s_k) \mid j \in \mathcal{C}\}$
4: $\mathcal{M}_i[id] \leftarrow$ top-n éléments les plus similaires de $\mathcal{M}_{i-1}[id] \cup \mathcal{N}_{id}$;
5: **foreach** $(j, s) \in \mathcal{N}_{id}$ **do**
6: $\mathcal{M}_i[j] \leftarrow$ top-n éléments les plus similaires de $\mathcal{M}_{i-1}[j] \cup \{(id, s)\}$;
7: **return** $\mathcal{M}_i$

4. Expérimentation

4.1. Protocole

Dans cette section, nous présentons une évaluation expérimentale de notre proposition. L'objectif principal de nos expériences est d'évaluer les performances de notre architecture de traitement en temps réel et sa capacité à traiter des données incrémentales. Notre évaluation vise à répondre aux questions de recherche suivantes :

QR1. Quelle est l'efficacité et la performance de notre proposition ?

QR2. Comment la taille des données d'entraînement incrémentales affecte-t-elle les résultats expérimentaux de l'apprentissage incrémental ?

QR3. Quelle est la quantité d'énergie consommée par ce processus ?

Datasets. Nous utilisons un dataset open-source créé à partir de données réelles et dédié à la tâche d'ER. Il est composé de données provenant de deux sources et décrit une liste de produits d'Amazon [2].

Métriques d'évaluation. Nous utilisons la précision, le rappel et le score F1 en tant que principales métriques de performance. Étant donné qu'il est difficile de définir la fin du processus lors du traitement des flux, nous adoptons une approche globale

2. https://zenodo.org/records/7930461

de l'exactitude après le traitement de toutes les données de la source B. Nous trions la liste des correspondances par ordre décroissant de similarité et sélectionnons $x(x \leq listLength)$ enregistrements dont la similarité se classe parmi les k premiers pour observer le rappel. En particulier, si une similarité correspond à plusieurs résultats, nous les prenons tous. Ensuite, nous transformons ces listes en paires de correspondances similaires sous forme de groupes de deux, de sorte que nous obtenons $N_{predicted}$ paires après suppression des doublons. Nous comparons ensuite le résultat avec les N_{truth} paires de la vérité terrain, où les enregistrements correspondent à la même entité, afin d'identifier le nombre de paires correctement prédites N_{common}.

Pour mesurer la consommation énergétique du programme, nous évaluons séparément la puissance du CPU et de la RAM en watts, ainsi que la consommation énergétique en joules.

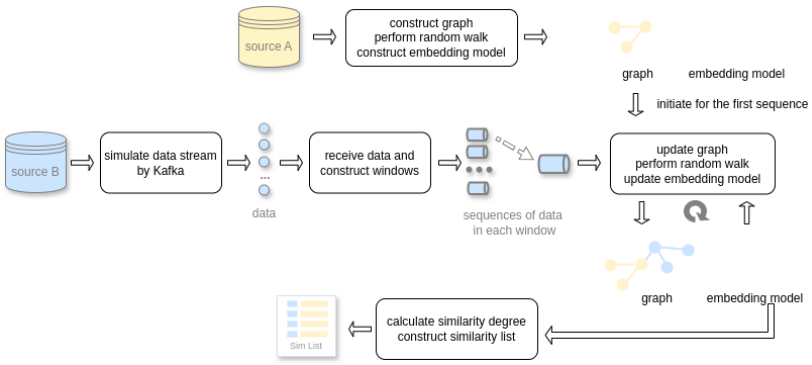


FIGURE 2. *Pipeline du processus d'implémentation*

Implémentation. Afin de simuler une application de traitement de données en flux, nous considérons les tuples de la source A comme un dataset local utilisé pour le pré-entraînement et envoyons les données de la source B vers un producteur Kafka de manière incrémentale, sous forme de flux de données à une fréquence fixe et régulière, sans concurrence. Pour chaque enregistrement de la source B, nous recherchons ses valeurs similaires dans la source A et construisons une séquence allant jusqu'à 10 éléments dans les deux directions (comme mentionné en Section 3.2) afin d'analyser dans quelle mesure le classement basé sur la similarité s'aligne avec les appariements corrects fournis par la vérité terrain. Dans cet article, nous utilisons directement les données relationnelles, et cet aspect ne sera donc pas approfondi davantage. L'ensemble du processus est illustré dans la Figure 2.

Sur le plan énergétique, nous utilisons Ecofloc (Alvarez-Valera *et al.*, 2024) pour mesurer la consommation énergétique. Ecofloc [3] est un WattMètre logiciel qui permet

3. https://github.com/labDomolandes/ecofloc

de mesurer l'énergie consommée par les processus en fonction de la charge qu'ils génèrent sur les principaux composants (CPU, GPU, RAM, NET, HD).

Les expériences ont été réalisées sur un ordinateur portable équipé d'un processeur 12×12th Gen Intel(R) Core(TM) i5-1245U avec 15,3 Go de RAM.

4.2. Analyse des résultats

4.2.1. Expériences sur l'efficacité

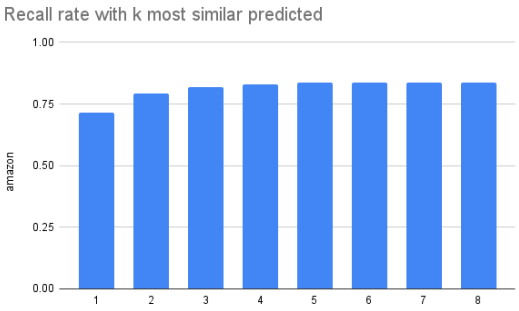


FIGURE 3. *Effet du nombre k de paires prédites les plus similaires sur le taux de rappel*

La Figure 3 illustre l'effet de l'ordre des valeurs les plus similaires sur le rappel, afin de démontrer l'efficacité de notre algorithme. Plus le rappel est élevé, plus le nombre de paires d'enregistrements correctement prédites est important. À mesure que la valeur de k augmente, nous observons que l'algorithme incrémental parvient à prédire avec succès la plupart des paires d'enregistrements similaires. De plus, la majorité des prédictions correctes sont concentrées parmi les enregistrements ayant des rangs de similarité relativement élevés.

L'ajustement de la valeur de k permet d'améliorer efficacement le rappel dans les tâches d'ER. En outre, au-delà d'un certain seuil, la longueur de la liste de similarité a un impact minimal sur la sensibilité des paires prédites. La plupart des paires ayant une faible similarité ne sont pas pertinentes, et privilégier les paires les plus similaires permet d'améliorer de manière significative à la fois l'efficacité et le rappel.

4.2.2. Effet de la taille des incréments sur les résultats

Dans l'apprentissage incrémental, la taille des nouvelles données ajoutées à chaque étape influence le processus d'entraînement du modèle. Des mises à jour trop fréquentes ou l'introduction d'un volume de données trop important en une seule fois peuvent déstabiliser le modèle. Afin d'analyser l'effet de la taille des données incrémentales sur la qualité des embeddings et sur la performance d'ER, nous avons fait

varier la proportion des données incrémentales. La taille des incréments est exprimée en pourcentage des données pré-entraînées. Par exemple, un incrément de 5% signifie que la quantité de données traitée à chaque étape incrémentale correspond à 5% du volume initial des données pré-entraînées.

TABLEAU 2. *Effect of increment size on results*

Taille de l'incrément	Temps d'exécution (s)	k	Précision (%)	Rappel (%)	Score F1 (%)
5.0%	326.50	1	34.7	59.3	43.5
		2	24.33	66.13	35.53
		3	19.93	67.33	31.00
10%	326.50	1	32.50	70.23	44.70
		2	22.33	78.67	34.67
		3	18.50	81.00	29.93
20%	332.00	1	34.40	71.33	46.40
		2	23.53	79.33	36.67
		3	20.00	81.80	32.00
30%	336.00	1	36.73	72.33	48.83
		2	25.87	79.67	38.90
		3	21.67	81.33	34.00
40%	338.00	1	34.50	70.00	46.33
		2	24.00	78.47	36.60
		3	20.00	81.23	32.40

Le Tableau 2 illustre l'évolution de la précision et du rappel en fonction de la taille de l'incrément. Le temps d'exécution et le rappel augmentent à mesure que la taille de l'incrément croît. On en déduit que, pour l'ER, bien que des mises à jour fréquentes du corpus textuel permettent un traitement plus rapide, elles ont un impact négatif sur les performances du modèle, réduisant sa capacité à identifier correctement les enregistrements similaires. En revanche, une taille d'incrément variant entre 20% et 40% n'a que peu d'effet sur les résultats.

Un rappel élevé indique que la majorité des appariements corrects sont identifiés, ce qui est crucial dans un contexte où d'importants volumes de données en flux continu arrivent rapidement. La réduction du nombre de paires potentielles constitue une base pour les étapes ultérieures qui visent à améliorer progressivement la précision, par exemple en passant d'une approche de traitement en flux à une méthode de traitement par lots plus efficace pour augmenter la précision.

4.2.3. Expériences sur la consommation d'énergie

Dans cette section, nous explorons l'étude de la consommation d'énergie dans les tâches d'apprentissage automatique ainsi que dans les tâches d'ERs.

TABLEAU 3. *Consommation d'énergie en fonction de la taille des incréments*

Incrément (% du pré-entraînement)	Temps d'exécution (s)	Consommation puissance CPU (W)	Consommation énergie CPU (J)	Consommation puissance RAM (W)	Consommation énergie RAM (J)
5%	326.50	1.23	401.76	2.18	711.28
10%	326.50	1.24	405.35	1.89	615.91
20%	332.00	0.97	323.48	2.03	673.21
30%	336.50	1.01	340.55	2.06	694.48
40%	338.00	1.05	354.11	2.08	704.64

Le tableau présente le temps d'exécution, la puissance et la consommation d'énergie calculée lors de l'exécution du programme pour différentes tailles de données incrémentales. À mesure que la quantité de données traitées en une seule session d'entraînement augmente, la consommation d'énergie tend d'abord à diminuer puis à augmenter. En effet, entraîner une grande quantité de données en une seule fois ou mettre à jour trop fréquemment consomme plus d'énergie que d'entraîner à plusieurs reprises une quantité modérée de données.

Bien que la chaleur dégagée par l'ordinateur lors de l'exécution du programme puisse influencer la consommation d'énergie (augmentation des besoins en refroidissement), les mesures effectuées offrent néanmoins une première estimation de l'impact énergétique de cette tâche. Il nous offre une orientation pour nos futures recherches sur l'optimisation de la consommation d'énergie.

De manière générale, il est évident que les embeddings de mots sont efficaces dans une approche incrémentale, permettant d'identifier la majorité des paires similaires correspondant à la même entité. De plus, certaines configurations, comme la taille des incréments, influencent les performances des embeddings de mots et, par conséquent, les résultats de la reconnaissance d'entités. Par ailleurs, ces configurations impactent également la consommation d'énergie durant l'exécution du code. Il est donc nécessaire d'explorer à la fois l'efficacité et la consommation énergétique afin de trouver un équilibre optimal entre les deux, permettant ainsi d'atteindre un objectif sans entraîner une consommation énergétique excessive.

5. Conclusion

Avec la demande croissante d'analyses de flux de données, il est essentiel d'intégrer efficacement les données en streaming avec les données stockées. Dans cette étude, nous avons abordé le défi d'ER dans l'intégration des données en adaptant les techniques d'embedding existantes, initialement conçues pour le traitement par lots, à un contexte de streaming. Nous avons proposé une solution dynamique basée sur un modèle d'embedding incrémental qui met continuellement à jour les représentations au fur et à mesure de l'arrivée des nouvelles données.

Plus précisément, nous avons construit un graphe hétérogène à partir de données relationnelles et l'avons mis à jour de manière incrémentale avec des séquences de données en continu. En exploitant ce graphe évolutif, nous avons effectué des random walks à partir des nouveaux nœuds ajoutés pour générer des phrases, qui ont ensuite été utilisées pour mettre à jour le modèle d'embedding. Notre approche a atteint un rappel allant jusqu'à 81.8%, démontrant ainsi son efficacité dans l'identification de correspondances similaires dans un contexte de streaming. Nous fournissons également des mesures quantitatives de la consommation d'énergie.

Cependant, notre méthode se concentre principalement sur la détection de paires d'enregistrements potentiellement similaires. Comme prochaine étape, nous avons pour objectif d'affiner nos résultats en exploitant les listes de similarité pour identifier

des correspondances plus définitives. Par exemple, en améliorant la méthode de calcul de similarité plutôt que d'utiliser uniquement la similarité cosinus. De plus, nous nous concentrerons sur l'identification des étapes du processus d'ERs en streaming qui consomment le plus d'énergie, dans le but d'optimiser cette partie de l'algorithme et de développer une approche plus efficace et durable.

Bibliographie

Almeida F., Xexéo G. (2023). *Word embeddings: A survey.*

Alvarez-Valera H. H., Maurice A., Ravat F., Song J., Roose P., Valles-Parlangeau N. (2024). Energy measurement system for data lake: An initial approach. In N. T. Nguyen *et al*. (Eds.), *Intelligent information and database systems*, vol. 14795, p. 1527. Singapore, Springer Nature Singapore.

Bahri M., Bifet A., Gama J., Gomes H. M., Maniu S. (2021, mars). Data stream analysis: Foundations, major tasks and tools. *WIREs Data Mining and Knowledge Discovery*, vol. 11, n° 3, p. e1405.

Barros C. D. T., Mendonça M. R. F., Vieira A. B., Ziviani A. (2021, novembre). A survey on embedding dynamic graphs. , vol. 55, n° 1.

Binette O., Steorts R. C. (2022, mars). (almost) all of entity resolution. *Science Advances*, vol. 8, n° 12, p. eabi8021.

Brunner U., Stockinger K. (2020). Entity matching with transformer architectures - a step forward in data integration. In *Proceedings of the 23rd international conference on extending database technology (edbt).*

Cappuzzo R., Papotti P., Thirumuruganathan S. (2020, juin). Creating embeddings of heterogeneous relational datasets for data integration tasks. In *Proceedings of the 2020 acm sigmod international conference on management of data*, p. 13351349. Portland OR USA, ACM.

Christophides V., Efthymiou V., Palpanas T., Papadakis G., Stefanidis K. (2020). An overview of end-to-end entity resolution for big data. *ACM Comput. Surv.*.

Do Nascimento D. C., Santos Pires C. E., Gomes Mestre D. (2018, mars). Heuristic-based approaches for speeding up incremental record linkage. *Journal of Systems and Software*, vol. 137, p. 335354.

Ebraheem M., Thirumuruganathan S., Joty S., Ouzzani M., Tang N. (2018, juillet). Distributed representations of tuples for entity resolution. *Proceedings of the VLDB Endowment*, vol. 11, n° 11, p. 14541467.

Gazzarri L., Herschel M. (2021, avril). End-to-end task based parallelization for entity resolution on dynamic data. In *2021 ieee 37th international conference on data engineering (icde)*, p. 12481259. Chania, Greece, IEEE.

Gazzarri L., Herschel M. (2023). Progressive entity resolution over incremental data. In *Proceedings of the 26th international conference on extending database technology (edbt).* OpenProceedings.org.

Genossar B., Shraga R., Gal A. (2023, mai). Flexer: Flexible entity resolution for multiple intents. *Proc. ACM Manag. Data*, vol. 1, n° 1.

Isah H., Abughofa T., Mahfuz S., Ajerla D., Zulkernine F., Khan S. (2019). A survey of distributed data stream processing frameworks. *IEEE Access*, vol. 7, p. 154300-154316.

Li Y., Li J., Suhara Y., Doan A., Tan W.-C. (2020, septembre). Deep entity matching with pre-trained language models. *Proceedings of the VLDB Endowment*, vol. 14, n° 1, p. 5060.

Maharana K., Mondal S., Nemade B. (2022). A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*, vol. 3, n° 1, p. 91-99. Consulté sur https://www.sciencedirect.com/science/article/pii/S2666285X22000565 (International Conference on Intelligent Engineering Approach(ICIEA-2022))

Marz N., Warren J. (2015). *Big data: Principles and best practices of scalable realtime data systems* (1st éd.). USA, Manning Publications Co.

Mikolov T., Sutskever I., Chen K., Corrado G., Dean J. (2013). Distributed representations of words and phrases and their compositionality. In, p. 31113119. Red Hook, NY, USA, Curran Associates Inc.

Mudgal S., Li H., Rekatsinas T., Doan A., Park Y., Krishnan G. *et al*. (2018, mai). Deep learning for entity matching: A design space exploration. In *Proceedings of the 2018 international conference on management of data*, p. 1934. Houston TX USA, ACM.

Papadakis G., Efthymiou V., Thanos E., Hassanzadeh O., Christen P. (2023, novembre). An analysis of one-to-one matching algorithms for entity resolution. *The VLDB Journal*, vol. 32, n° 6, p. 13691400.

Pennington J., Socher R., Manning C. D. (2014). Glove: Global vectors for word representation. In *Empirical methods in natural language processing (emnlp)*, p. 1532–1543.

Saeedi A., Peukert E., Rahm E. (2020). Incremental multi-source entity resolution for knowledge graph completion. In *The semantic web*, vol. 12123, p. 393408. Cham, Springer International Publishing.

Simonini G., Zecchini L., Bergamaschi S., Naumann F. (2022, mars). Entity resolution on-demand. *Proceedings of the VLDB Endowment*, vol. 15, n° 7, p. 15061518.

Tu J., Fan J., Tang N., Wang P., Li G., Du X. *et al*. (2023, mai). Unicorn: A unified multi-tasking model for supporting matching tasks in data integration. *Proceedings of the ACM on Management of Data*, vol. 1, n° 1, p. 126.

Wang S., Zhou W., Jiang C. (2020, mars). A survey of word embeddings based on deep learning. *Computing*, vol. 102, n° 3, p. 717740.

Zeng X., Wang P., Mao Y., Chen L., Liu X., Gao Y. (2024). Multiem: Efficient and effective unsupervised multi-table entity matching. In *2024 ieee 40th international conference on data engineering (icde)*, p. 3421-3434.

Gouvernance des données

Vers un cadre conceptuel

Jacky AKOKA[1], Isabelle COMYN-WATTIAU[2]

1 Laboratoire CEDRIC-CNAM, 2 Rue Conté, 75003 PARIS
2. ESSEC Business School, 3 Av. B. Hirsch, 95021 CERGY Cedex

RESUME. *La gouvernance des données est une activité essentielle pour les organisations qui cherchent à exploiter les données comme un atout stratégique. Son objectif est de maximiser la valeur tout en minimisant les coûts et les risques. Dans cet article, nous présentons un cadre conceptuel pour la gouvernance des données offrant une vue holistique enrichissant les cadres et modèles académiques et professionnels existants. En utilisant des techniques bibliométriques, nous analysons la littérature existante afin d'identifier les éléments clés de la gouvernance des données, notamment sa structure intellectuelle, ses thèmes de recherche et les articles les plus influents qui en forment la colonne vertébrale. Dans un deuxième temps, nous proposons un cadre conceptuel enrichi fondé sur la théorie des systèmes. Ce cadre englobe cinq dimensions primordiales : le but, la structure, les activités, l'environnement et le résultat. Il permet également de prendre en considération l'interaction entre ces dimensions. Afin d'illustrer ce cadre conceptuel, nous décrivons comment celui-ci a permis de structurer les questions d'un baromètre dédié à l'évaluation de la maturité de la gouvernance des données dans les organisations. Nous présentons ensuite quelques cas d'usage, puis nous discutons des implications pour les chercheurs et les praticiens.*

ABSTRACT. *Data governance constitutes an indispensable activity for organizations intent on leveraging data as a strategic asset. Its objective is to maximize value while minimizing costs and risks. In this article, we present a conceptual framework for data governance that offers a holistic view, thereby enriching existing academic and professional frameworks and models. Using bibliometric techniques, we analyze the extant literature to identify the key elements of data governance, including its intellectual structure, research themes, and the most influential articles that form its backbone. Secondly, we propose an enriched conceptual framework based on systems theory. This framework encompasses five overarching dimensions: goal, structure, activities, environment, and outcome. It also considers the interaction between these dimensions. To illustrate this conceptual framework, we describe how it was used to structure the questions of a barometer dedicated to assessing data governance maturity in organizations. We then present a few use cases and discuss the implications for researchers and practitioners.*

MOTS-CLES : *gouvernance de données, cadre conceptuel, valeur, risque, coût, étude bibliométrique.*

KEYWORDS: *data governance, conceptual framework, value, risk, cost, bibliometric study.*

1. Introduction

Les données sont un atout stratégique et une ressource pour les organisations, car elles conditionnent la prise de décision. La gouvernance des données permet la gestion des données de manière efficace pour améliorer la performance de l'organisation tout en respectant les obligations réglementaires et en minimisant les coûts et les risques. La gouvernance des données désigne l'ensemble des processus, politiques et outils qui permettent de déterminer qui est responsable de la prise de décision d'une organisation au sujet de ses données (Khatri et Brown, 2010). En cela, ils distinguent la gouvernance des données de leur management, lequel consiste à prendre et mettre en œuvre ces décisions. (Smallwood, 2019) la définit comme un ensemble de structures, politiques, procédures, processus et technologies utilisés pour collecter, organiser, utiliser et sécuriser les données. Plusieurs revues de la littérature sur la gouvernance des données ont été publiées (Nguyen, 2016 ; Abraham *et al*., 2019 ; Merkus *et al*., 2019). La gouvernance des données n'est pas simplement une question de technologie ou de management des données ; il s'agit fondamentalement d'établir les règles et les responsabilités relatives au management des données. Il manque, à ce jour, une définition acceptée par tous, tant chercheurs que praticiens. De plus, de nombreux cadres de référence (« frameworks ») ou modèles ont été proposés sans qu'aucun ne fasse consensus. Notre recherche est ainsi guidée par les questions suivantes : QR1 : *Quelle est la structure intellectuelle, les thèmes de recherche et les articles influents dans le domaine de la gouvernance des données ?* QR2 : *Quel cadre théorique peut-on proposer pour structurer les composants de ce domaine ?*

Pour répondre à ces questions, nous effectuons une analyse rétrospective complète de la littérature relative à la gouvernance des données dans la base de données Scopus (QR1). Pour ce faire, trois méthodes d'analyse bibliométrique quantitative sont employées : l'analyse des co-citations (CCA), l'analyse du couplage bibliographique (BCA) et l'analyse des chemins principaux (MPA). Grâce à cette étude bibliométrique, nous identifions les dimensions les plus saillantes et les lacunes, ce qui nous permet de développer un nouveau cadre conceptuel (QR2). Notre cadre n'est pas destiné à remplacer les cadres existants, mais plutôt à fournir une carte conceptuelle qui montre comment les différents éléments s'assemblent pour définir la gouvernance des données d'une manière multidimensionnelle.

Dans la section qui suit, nous présentons un état de l'art structuré autour d'une étude bibliométrique. Le cadre conceptuel fondé sur la théorie des systèmes est décrit dans la section 3. Dans la section 4, nous illustrons l'utilisation de ce cadre pour générer la structure d'un baromètre dédié à l'étude de la maturité des organisations en gouvernance des données et décrivons quelques cas d'usages. La section 5 conclut et suggère des voies de recherche future.

2. Étude bibliométrique de la gouvernance des données

L'objet principal de cette section est de dresser un état de l'art de la littérature scientifique sur la gouvernance des données. Au-delà, nous visons à cartographier

cette littérature à l'aide de techniques bibliométriques. A cette fin, nous utilisons trois techniques d'analyse des citations, après une étude descriptive du domaine. Dans la suite, nous présentons le dispositif méthodologique fondé sur trois étapes : la constitution du jeu de données, l'analyse descriptive et l'analyse par les citations.

2.1. Constitution du jeu de données

La constitution du jeu de données requiert le choix d'une base de données bibliographique, la construction d'une requête, l'extraction des résultats et le nettoyage des données. Dans cette recherche, nous avons soumis à la base de données bibliographiques Scopus, réputée pour sa pertinence dans le domaine de l'informatique et des systèmes d'information, la requête ci-dessous :

```
   TITLE-ABS-KEY("data governance") AND ( EXCLUDE ( DOCTYPE,"ch" ) OR
EXCLUDE ( DOCTYPE,"bk" ) OR EXCLUDE ( DOCTYPE,"no" ) OR EXCLUDE (
DOCTYPE,"ed" ) OR EXCLUDE ( DOCTYPE,"sh" ) OR EXCLUDE ( DOCTYPE,"er" )
OR EXCLUDE ( DOCTYPE,"le" ) OR EXCLUDE ( DOCTYPE,"tb" ) ) AND ( LIMIT-
TO ( LANGUAGE,"English" ) )
```

La chaîne de recherche choisie est « data governance » limitée aux titres, résumés et mots-clés des publications en anglais dans des revues ou des conférences, ce qui nous retourne 2701 documents. Certains auteurs utilisent de manière interchangeable l'expression « gouvernance de l'information ». L'ajout de cette chaine dans la requête n'a pas apporté de modifications significatives pertinentes. Après exportation des données, à l'aide du logiciel CRExplorer (Thor *et al*., 2021), nous avons procédé à la déduplication des références. En effet, Scopus ne produit pas de documents en double mais, dans ces documents, de nombreuses références sont mal codées, rendant difficile leur rapprochement par les logiciels d'analyse bibliométrique. CRExplorer dispose d'un outil de rapprochement et d'harmonisation de ces références. Un nettoyage « manuel » a aussi été nécessaire pour éliminer quelques références incomplètes qui perturbaient les analyses, en l'occurrence une référence, ce qui nous a conduit à un jeu de données de 2700 articles.

2.2. Analyse descriptive

Avant d'utiliser les techniques d'analyse des citations, nous avons procédé à une analyse descriptive du jeu de données résultant de l'étape précédente. La distribution des articles obtenus est représentée à la figure 1. Le terme « gouvernance des données » apparait pour la première fois en 2005 et le phénomène reste embryonnaire jusqu'en 2011 environ.

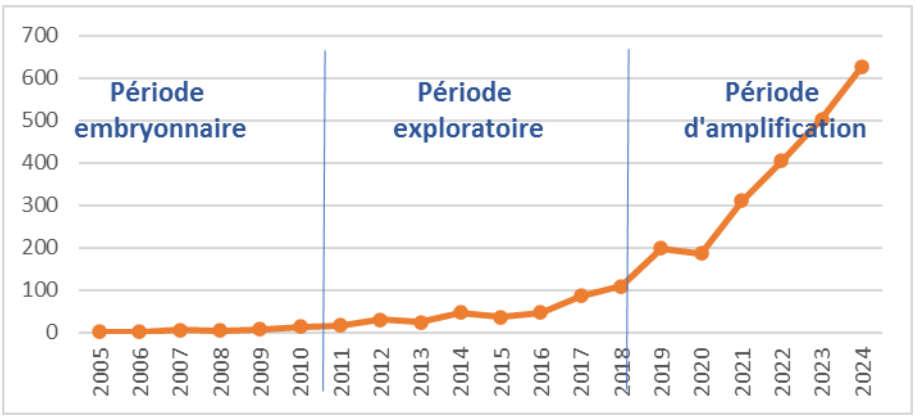


Figure 1. Nombre de publications par an

La période 2011 à 2018, qu'on peut qualifier d'exploratoire, révèle une augmentation régulière passant de 16 publications en 2011 à 109 en 2018. La période de 2018 à 2024, ou période d'amplification, voit le nombre de publications augmenter beaucoup plus rapidement jusqu'à atteindre 626 en la seule année 2024. Le terme de gouvernance de données est ainsi adopté progressivement. Le domaine de la gouvernance des données fait ainsi l'objet d'une recherche plus abondante. Scopus fournit un ensemble de tableaux et de graphiques décrivant certains traits saillants du jeu de données renvoyé par la requête. Nous avons sélectionné les plus pertinents. Ainsi, en ce qui concerne les auteurs les plus prolifiques (Fig. 2), Rob Brennan se détache avec plus de 20 publications centrées autour de la valeur de la donnée, sa qualité et la mesure de sa valeur (Brennan *et al*., 2018). Les auteurs ne sont sélectionnés ici que sur leur nombre de publications dans le jeu de données, c'est-à-dire dans les articles « étiquetés » gouvernance de données.

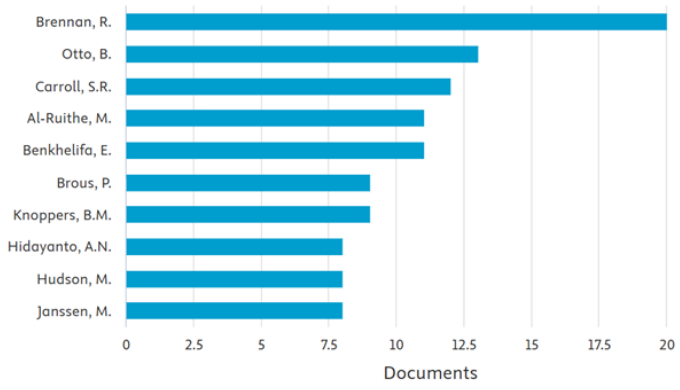


Figure 2. Auteurs les plus prolifiques

La distribution des documents par type montre deux pôles équilibrés entre les articles publiés dans des revues et ceux parus dans les actes de conférence (Fig. 3).

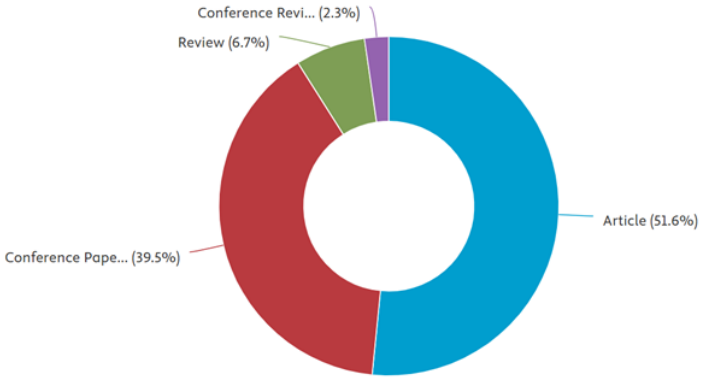


Figure 3. Documents par type

La recherche relative à la gouvernance des données est multidisciplinaire, avec toutefois près de 30% pour l'informatique (« computer science ») (Fig. 4). A noter qu'un domaine d'application, la médecine, émerge nettement (près de 9% des publications y sont rattachées).

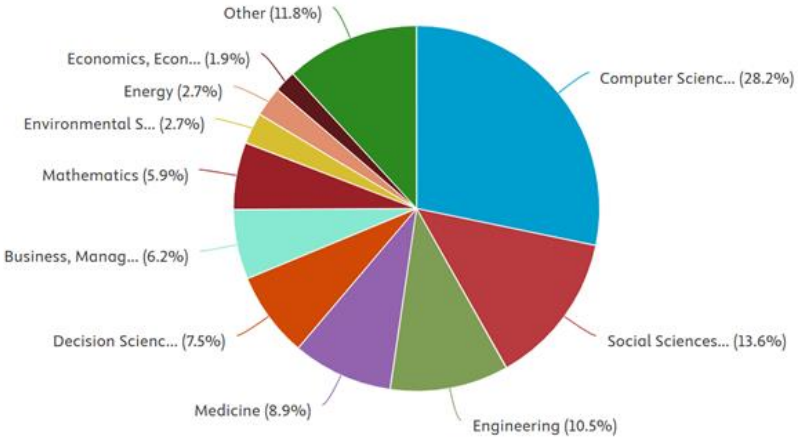


Figure 4. Documents par discipline

En résumé, cette analyse permet d'identifier une recherche en croissance en gouvernance des données, avec quelques auteurs très présents et des sujets comme la qualité de la donnée, sa confidentialité, sa valeur et donc le besoin de gouvernance.

2.2. Analyse par les citations

Les publications extraites sur la gouvernance des données sont trop nombreuses pour en faire une analyse individuelle. Dès lors, le recours aux techniques bibliométriques est judicieux. Dans cette section, nous présentons l'application des techniques de co-citation (« co-citation analysis ») pour identifier la structure

intellectuelle du domaine. Puis le couplage bibliographique (« bibliographic coupling analysis ») nous permet de repérer les thèmes principaux. Enfin, au moyen du calcul de chemins principaux (« main path analysis »), nous mettons en évidence les articles les plus influents et les liens de citation qui les relient.

2.2.1. Analyse des co-citations (CCA)

L'analyse des co-citations consiste à construire des classes (clusters) d'articles (ou références) cités par un ensemble d'articles citants. Plus deux références sont citées ensemble, plus elles sont proches et susceptibles d'appartenir au même cluster. Le logiciel utilisé ici est Vosviewer (pour Visualization of Similarities) met en œuvre une similarité appelée force d'association (association strength) pour répartir les articles dans le plan (Van Eck et Waltman, 2013). A l'aide de Vosviewer, nous avons catégorisé trois domaines de références regroupant les 28 articles cités plus de 30 fois dans notre ensemble de publications (Fig. 5). Le choix du seuil de 30 est empirique et dicté par le souci de ne retenir que des articles influents et d'obtenir une visualisation de taille raisonnable. La première catégorie (14 articles en rouge) regroupe les publications proposant des modèles et des cadres conceptuels (« frameworks ») sur le thème de l'organisation, les rôles, les responsabilités et la prise de décision relatifs à la gouvernance des données. Ainsi, Khatri et Brown (2010) différencient la gouvernance des données de son management et proposent un cadre distinguant les cinq domaines de décision relative à la gouvernance des données : les principes qui gouvernent l'usage des données, la qualité des données, les méta-données, l'accès aux données et le cycle de vie de la donnée. L'article de Weber *et al*. (2009) introduit un modèle matriciel de gouvernance de la donnée qui croise les rôles et les activités. Un autre article séminal décrit une étude empirique de la gouvernance de l'information concernant trente organisations (Tallon *et al*., 2013). Ils en déduisent un modèle composé d'antécédents, facilitateurs ou inhibiteurs, des mécanismes structurels, procéduraux ou relationnels de la gouvernance de l'information avec des résultats en termes de performance et de mitigation des risques. A noter qu'il s'agit de la gouvernance de l'information et non de la gouvernance des données, bien que ces deux termes soient souvent utilisés de manière interchangeable.

La seconde catégorie (9 articles en vert) regroupe les publications proposant des cadres conceptuels de gouvernance de données intégrant davantage de dimensions. Ces articles paraissent dans des revues relevant de domaines très différents (informatique, juridique, politique, psychologique, managérial). Le plus cité est (Abraham *et al*., 2019) qui étudie 145 sources afin d'identifier les composants de la gouvernance des données selon six dimensions : le domaine concerné (la qualité des données, la sécurité, les méta-données, raffinant les cinq domaines de Khatri et Brown), le périmètre organisationnel concerné (un département, toute l'organisation, l'inter-organisationnel), le type de données concerné (traditionnelles ou massives), les mécanismes de gouvernance (structurels, procéduraux ou relationnels), les antécédents tant internes qu'externes et les conséquences (performances et risques). Pour les données scientifiques, Wilkinson *et al*. (2016) proposent les quatre principes FAIR (pour facile à trouver, accessible, interopérable et réutilisable) pour

un bon management des données et une bonne intendance (stewardship). Janssen *et al.* (2020) en proposent treize à appliquer aux systèmes algorithmiques des big data, dont beaucoup reprennent les règles relatives aux données personnelles (par exemple la minimisation de l'accès aux données ou encore la nécessité d'informer).

La troisième catégorie (5 articles en bleu) est dominée par des analyses de la littérature naissante sur la gouvernance de données. Ainsi, Alhassan *et al*. (2016) proposent une analyse de fréquence des activités de gouvernance de données selon l'action (définir, mettre en œuvre et surveiller) effectuée dans un espace (standards liés aux données, exigences liées aux données, stratégie en matière de données, politique des données, etc.) pour un des cinq domaines définis par Khatri et Brown (2010). Al-Ruithe *et al*. (2019) présentent une analyse systématique de la littérature qui compare la gouvernance des données en général à celle liée au cloud. Six dimensions sont ainsi comparées : la politique de gouvernance de la donnée, son administration, sa structure organisationnelle, la dimension technologique, la dimension réglementaire et ses outils de mesure et de surveillance.

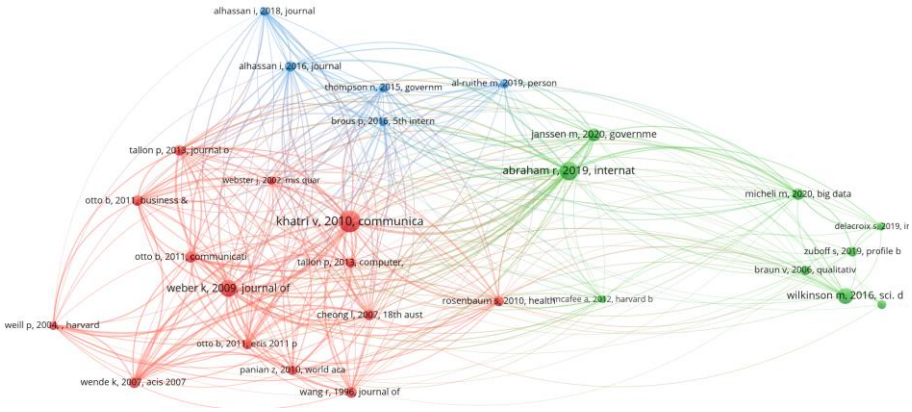


Figure 5. Analyse des co-citations

En conclusion, cette analyse nous a permis d'identifier la structure intellectuelle de la gouvernance des données, une première composante principalement orientée vers la dimension organisationnelle du sujet, une seconde élargissant la dimension à des cadres conceptuels plus complets et l'appliquant à des domaines variées, la troisième composante s'attachant à l'étude de la littérature naissante sur le sujet.

2.2.2. Analyse des couplages bibliographiques (BCA)

L'analyse des couplages bibliographiques consiste à construire des classes d'articles considérés proches parce qu'ils citent un même ensemble de références Plus cet ensemble commun est grand, plus ils sont susceptibles d'appartenir au même cluster. Elle permet de rapprocher les articles travaillant sur des thématiques proches. D'autre part, elle complète l'analyse des co-citations puisqu'elle permet

d'intégrer aussi des articles récents, lesquels ne sont pas encore assez cités pour apparaître dans l'analyse des co-citations.

L'analyse des couplages bibliographiques produite par Vosviewer catégorise en 6 classes les 102 articles cités plus de 50 fois et connectés (Fig. 6). Le seuil de 50 est défini de façon empirique pour retenir les articles les plus influents. Vosviewer permet aussi de choisir le nombre de classes souhaité en modifiant la valeur de la résolution. Nous avons effectué ce calibrage en examinant, par essai-erreur, le contenu des classes et surtout leur interprétabilité.

La première classe (35 nœuds en rouge) regroupe des articles étudiant les problématiques de gouvernance de données dans le domaine de l'apprentissage automatique, de la santé et de la « blockchain ». La deuxième classe (28 nœuds en vert) contient de nombreux articles déjà mentionnés dans l'analyse des co-citations. Elle regroupe donc les articles fondateurs sur la gouvernance de données, notamment les différents cadres conceptuels. La troisième classe (16 nœuds en bleu) traite de la gouvernance de données appliquée à différents domaines dont les « smart cities », l'aspect juridique, les plates-formes de données, le « big data ». La quatrième classe (14 nœuds en jaune) est aussi caractérisée par le big data mais aussi d'autres domaines d'application, comme l'agriculture. La cinquième classe (7 nœuds en violet) cible le domaine de l'industrie 4.0 et celui des « smart cities ». Enfin, la sixième classe (2 nœuds en bleu clair) porte sur l'interdépendance entre bases de données et intelligence artificielle. A noter que le thème des « smart cities » commun aux classes 3 et 5 est traité sur un plan organisationnel dans la première et plus technologique dans la seconde. En conclusion, on a une seule classe (la deuxième) qui concentre les articles sur les fondamentaux de la gouvernance de données tandis que les autres les appliquent à différents domaines.

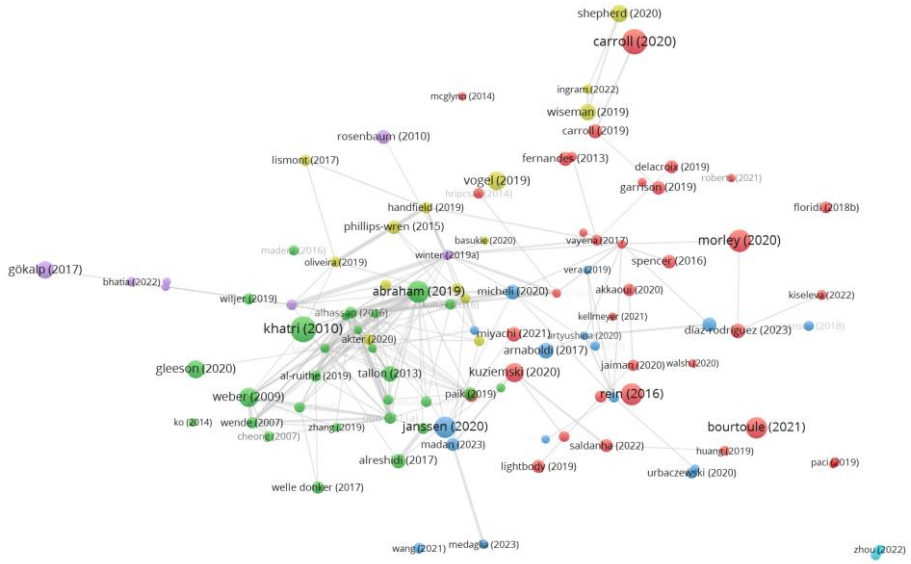


Figure 6. Analyse des couplages bibliographiques

Ainsi, deux thèmes principaux émergent. Le premier est lié aux fondements de la gouvernance de données. Le second, regroupant le reste des classes, est dédié aux applications dans différents domaines.

2.2.3. Analyse des chemins principaux (MPA)

L'analyse des chemins principaux s'effectue sur le graphe de citations entre les articles du jeu de données. Elle consiste à rechercher les trajectoires principales de ce graphe. La métrique utilisée ici est SPLC (Search Path Link Count) qui compte le nombre de fois où un arc est traversé si l'on parcourt tous les chemins possibles depuis tous les ancêtres du nœud d'origine vers tous les puits du graphe. A l'aide de cette métrique, on peut ensuite calculer les chemins les plus influents au moyen de différents algorithmes. Le logiciel Pajek propose plusieurs algorithmes. Celui utilisé ici est le global key-route (ici 10 key-routes) qui consiste à retenir les dix arcs ayant la valeur de SPLC la plus importante et, pour chacun de ces arcs, à calculer les chemins passant par cet arc. On obtient le graphe de la figure 7 qui représente une vision synthétique de la structuration du domaine de la gouvernance des données et les différents canaux de diffusion de la connaissance. Parmi ces chemins, le plus impactant est en rouge sur la figure.

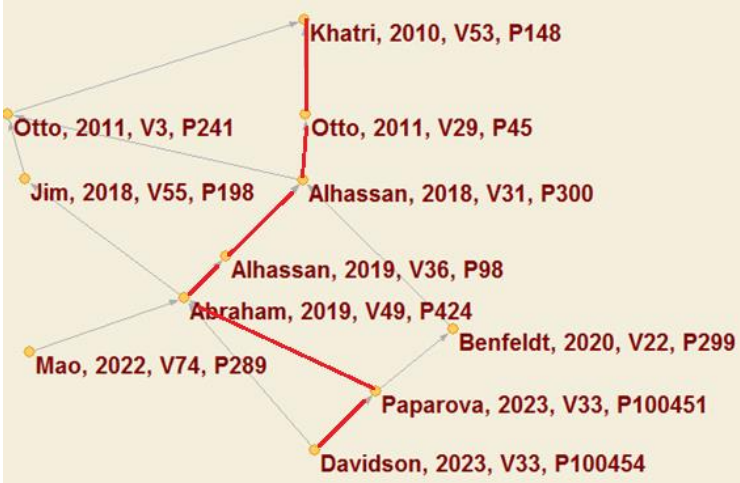


Figure 7. Analyse des chemins principaux de la gouvernance de données

Le nœud racine est l'article de Khatri et Brown (2010) déjà mentionné dans la CCA. Otto (2011) compare l'organisation de la gouvernance de données dans deux entreprises de télécommunications selon trois dimensions (le but, la structure et les processus). Alhassan *et al.* (2018) comparent la littérature scientifique et la presse professionnelle selon un cadre qui croise les cinq domaines de décision définis par Khatri et Brown (2010) et les trois actions Définir, Mettre en œuvre et Surveiller. Les mêmes auteurs identifient aussi les facteurs clés de succès de la gouvernance des données au travers de la littérature (Alhassan *et al.*, 2019). Abraham *et al.*

(2019) proposent un cadre de la gouvernance des données qui combine 3 dimensions (types de données, niveau organisationnel, processus de mise en œuvre) pour analyser les mécanismes de gouvernance, leurs antécédents et les résultats de la gouvernance des données. Paparova *et al*. (2023) introduisent le concept d'espace de gouvernance de données qui combine une logique verticale et une logique horizontale en matière de rôles et responsabilités. Cet article fait partie d'un numéro spécial sur la gouvernance des données dont l'éditorial porte sur les défis de la gouvernance des données à l'ère de l'innovation digitale (Davidson *et al*., 2023). Ce chemin principal constitue une sorte de colonne vertébrale (« backbone ») de diffusion de la connaissance du domaine naissant de la gouvernance de données

Pour des raisons d'espace, nous n'avons pas pu effectuer une analyse par période. Les différentes analyses menées font ressortir le concept de but de la gouvernance de donnée et celui de périmètre. Certains articles mettent l'accent sur les activités liées à la gouvernance des données, par exemple les efforts de mise en conformité. D'autres pointent sa structure qui comprend les rôles et les responsabilités. D'autres encore identifient des éléments facilitateurs ou des facteurs clés de succès. Enfin, certains auteurs structurent tout ou partie de ces éléments en un cadre conceptuel. A notre connaissance, aucun de ces cadres n'utilise un fondement théorique ni ne considère une vue holistique de la gouvernance de données. C'est un tel cadre que nous proposons dans la section suivante.

3. Vers un cadre de référence pour la gouvernance des données

Notre cadre de référence est fondé sur la théorie des systèmes (Skyttner, 1996). Comme nous l'avons établi dans (Akoka et Comyn-Wattiau, 2019), la gouvernance des données est un artefact de conception que l'organisation développe et met en œuvre pour atteindre ses objectifs. Simon (1996) caractérise les artefacts en termes de fonctions (*activités*), *d'objectifs* et d'adaptation (*évolution*). Il distingue également la *structure* de l'artefact de *l'environnement* dans lequel il opère. De plus, la gouvernance des données est au cœur d'un processus de pilotage dont elle ne peut être définie indépendamment. Son objet est la prise de décision, à la fois pour fixer les orientations de l'action (les *objectifs*), et pour définir et ajuster le cadre de fonctionnement (*structure* et *environnement*) et les *activités* correspondantes. En fixant les objectifs, les modalités et les règles du jeu des activités de management de données, la gouvernance des données est également amenée à les faire *évoluer* si nécessaire, en fonction des *résultats* obtenus : c'est là son rôle de pilote principal de la donnée. La gouvernance des données trouve sa signification dans la mise en relation des éléments qui la composent (Le Moigne, 1994 ; Skyttner, 1996). Ces dimensions forment un ensemble indissociable donnant lieu à une évaluation des *résultats* bénéficiant d'une *rétroaction*. À ce titre, la gouvernance des données possède toutes les caractéristiques d'un système. Nous présentons ci-dessous les composants de ce système et leur interaction (Fig. 8).

La dimension **But** de notre système comporte deux sous-dimensions. La première sous-dimension se rapporte à l'objectif de la gouvernance des données.

Bien sûr son but est dicté par l'alignement avec la stratégie mais il reste celui de maximiser sa valeur en minimisant les coûts et les risques (Bennett, 2017).

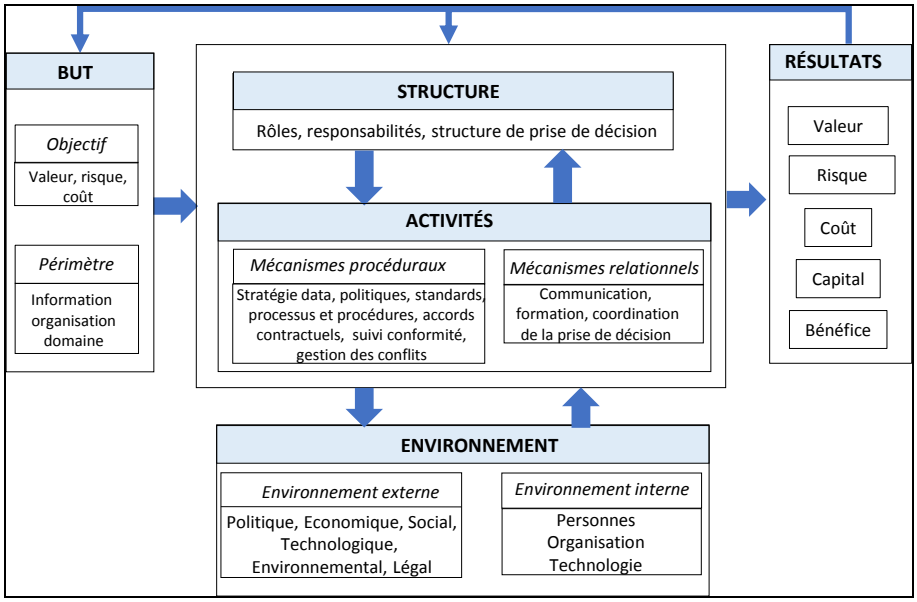


Figure 8. Le cadre de référence

La deuxième sous-dimension précise le périmètre auquel s'applique la gouvernance. Trois composants définissent ce périmètre, à savoir le périmètre « donnée », celui de « l'organisation » et celui du « domaine », tel que suggéré par Abraham *et al*. (2019). Dans notre cadre, le périmètre « donnée » intègre les données traditionnelles (structurées), les données massives (semi-structurées ou non structurées) mais aussi les données synthétiques générées par l'intelligence artificielle. Le périmètre « organisation » englobe les niveaux intra- et inter-organisationnels, mais peut aussi ne s'appliquer qu'à une entité de l'organisation. Enfin, le périmètre « domaine » est décrit par sept éléments, à savoir la qualité des données, la sécurité, l'architecture, le cycle de vie, les métadonnées, le stockage et l'infrastructure, ainsi que la gestion des documents et du contenu. La deuxième dimension a trait à la **Structure**, en liaison avec les mécanismes structurels proposés dans (Tallon *et al*., 2013). Ces mécanismes englobent les rôles et les responsabilités ainsi que l'attribution du pouvoir de décision. La troisième dimension, appelée **Activités**, identifie les mécanismes procéduraux et les mécanismes relationnels de la gouvernance tels qu'ils sont proposés dans (Tallon *et al*., 2013). Les premiers englobent la stratégie d'information, les politiques, les règles, les normes, les processus et les procédures, les accords contractuels, le contrôle de la conformité et la gestion des problèmes. Les mécanismes relationnels font référence à la communication, à la formation et à la coordination de la prise de décision. La quatrième dimension ou **Environnement** du système est composée de facteurs externes et internes à l'organisation et ayant un impact sur sa gouvernance des

données. Tallon *et al.* (2013) identifient un facteur externe, celui de la réglementation. Nous proposons d'enrichir cet ensemble de facteurs externes à l'aide du modèle PESTLE (Rastogi *et al*., 2016), incluant tant l'impact de la politique, que l'économie et le social, mais aussi les éléments technologiques, légaux et environnementaux. Comme le montrent Tallon *et al*. (2013), l'environnement est également composé de différents facteurs internes à l'organisation, ayant un impact potentiel sur le système de gouvernance de la donnée. Plusieurs auteurs mentionnent une liste d'éléments (culture, stratégie informatique, soutien de la direction générale, etc.) Nous proposons de structurer ces facteurs en trois catégories : les personnes, l'organisation et la technologie (Hevner *et al*., 2004). La cinquième dimension **Evolution** décrit les changements du système de gouvernance des données au fil du temps. Elle n'apparait pas dans la figure 8 qui représente la gouvernance à un moment donné. Ces changements peuvent affecter toutes les autres dimensions. Elle n'est pas décrite dans les cadres de référence de la littérature, autres que les modèles de maturité. Elle constitue une dimension fondamentale permettant aux entreprises de construire et d'alimenter ces modèles de maturité. Enfin, le système comprend une dimension de performance (valeur, risque, coût, bénéfice, capital investi) qui permet d'exercer la boucle de rétroaction pour faire évoluer le système.

En conclusion, en mobilisant la théorie des systèmes comme fondement de ce cadre conceptuel, nous offrons une vue holistique de la gouvernance des données composée de cinq dimensions, permettant d'identifier les interactions existant entre elles. Notre contribution consolide les différents axes de la recherche en matière de gouvernance des données en un modèle unique enrichi avec les composants d'autres domaines (Hevner *et al*., 2004 ; Rastogi *et al*., 2016 ; Bennett, 2017).

Il existe des cadres de référence proposés par les praticiens[1] (DAMA-DMBoK, COBIT, DGI, SAS, BCG). Malgré plusieurs points communs (principes, rôles et responsabilités organisationnels, politiques, règles et normes, processus de base, etc.), chaque cadre a un centre d'intérêt et son propre champ d'application : certains se concentrent sur la gouvernance stratégique, d'autres sur le management en termes de pratiques détaillées de gestion des données, d'autres encore sur des objectifs spécifiques (qualité, conformité, valeur). Notre cadre conceptuel offre une vue globale qui peut être déclinée par niveau et par domaine d'application, par exemple données massives et intelligence artificielle, secteur public et villes intelligentes.

4. Applications du cadre de référence

Ce cadre a été utilisé pour générer la structure d'un baromètre d'évaluation de la maturité des entreprises et organisations françaises en matière de gouvernance des données, appelé MetraData[2]. Cette enquête a été menée auprès de 150 entités. Nous

[1] https://www.kellton.com/kellton-tech-blog/popular-data-governance-frameworks#:~:text=,activities%20like%20metadata%20and%20architecture
[2] https://drive.google.com/file/d/1Gx3-mE0eRdizmIw4npObeXKBLm1daX3t/view

avons généré un questionnaire en ligne comprenant environ 40 questions auprès de dirigeants d'entreprise, directeurs des systèmes d'information, directeurs de données et directeurs fonctionnels dans tous les secteurs (privé, public, parapublic, association), tous les domaines d'activité (assurance, technologie, banque, etc.) et toutes les tailles d'entreprises. On a recueilli la perception du concept de gouvernance des données, son opérationnalisation dans l'organisation, ses acteurs-clés, le contexte et les plans d'évolution. Le baromètre met en lumière l'importance des facteurs organisationnels et humains permettant d'améliorer la gouvernance des données. La première colonne du tableau 1 reprend chaque dimension du cadre de référence. La seconde colonne contient un exemple de question fondée sur cette dimension. La troisième colonne fournit les modalités de réponse proposées

Tableau 1. Correspondance entre le cadre de référence et le baromètre MetraData

Dimension/ Sous-dimension	**Question**	**Réponses**
But/objectif	Quels sont les objectifs principaux de la gouvernance des données au sein de votre organisation ?	Maximiser la valeur métier tirée des données Assurer la conformité aux lois et réglementations Minimiser les risques liés aux données Minimiser les coûts de gestion des données
But/périmètre/ donnée	Quels sont les principaux périmètres d'intervention de la gouvernance des données dans votre organisation ?	Bases de données classiques, documents, "big data", emails, sites web, supports numériques (réseaux sociaux), supports papier, données synthétiques de l'IA
Structure/ Rôles et responsabilités	Quels sont les principaux rôles et responsabilités de la gouvernance des données identifiés dans votre organisation ?	Gestionnaire de données, architecte de données, propriétaire de données, directeur de données, analyste des données, délégué à la protection des données
Structure/Prise de décision	Quelles instances de votre organisation ont en charge tout ou partie de la gouvernance des données ?	Comité stratégique, comité risques et sécurité, comité de direction, comité de pilotage dédié, aucun
Activités/Mécanismes procéduraux	Les mécanismes ci-dessous sont-ils mis en place dans votre organisation pour opérationnaliser les activités de gouvernance des données ?	Gestion de la conformité Preuve et signature électronique Sécurité des données Architecture d'entreprise et modélis. données et flux Architecture des données Gestion de la qualité Définition des rôles liés aux données Investigation informatique (e-discovery) Archivage électronique Gestion des méta-données et catalogage Gestion des données maîtres et de référence Analyse des risques informationnels Gestion des documents et contenus
Activités/Mécanismes relationnels	Comment qualifiez-vous le dialogue entre les respon-sables de la gouvernance des données et les métiers dans votre organisation ?	Excellent, bon, améliorable, tendu, inexistant
Environnement/Externe	Quel a été l'élément déclencheur qui a suscité une préoccupation pour la gouvernance des données ?	Fusion/acquisition Cyber-attaque Fuite de donnée Nouvelle réglementation (e.g. RGPD), etc.

Environnement/ Interne/ Technologie	Quels sont les outils spécifiques que vous avez mis en place dans votre organisation afin d'opérationnaliser les activités de gouvernance des données ?	Master Data Management Mise en qualité des données Gestion documentaire Modélisation et architecture de données Classification et catalogage de données Registre de traitement RGPD, etc.
Environnement/ Interne/ Personnes et Organisation	Quels sont les éléments qui freinent ou pourraient freiner l'implantation de la gouvernance des données dans votre organisation ?	Manque de moyens humains Résistance au changement Culture d'entreprise éloignée de ces préoccupations Manque de dialogue métiers et équipes gouv. donnée Manque d'engagement de la direction Manque de dialogue entre la DSI et les équipes gouvernance donnée La data n'est pas un enjeu clé
Résultats	Quels éléments de la gouvernance des données sont suivis dans le reporting régulier de l'organisation ?	Avancement des projets Budget Niveau de satisfaction des services data Gestion des compétences

D’autres applications sont à l’étude. Premièrement, le cadre fournit un modèle théorique pour les recherches futures sur la gouvernance des données. De nouvelles questions de recherche peuvent émerger en croisant les principales dimensions. À titre d'illustration, l'étude de l'évolution de la gouvernance des données dans une organisation ainsi qu'au niveau sectoriel est un domaine de recherche prometteur. Un cadre offre de nombreux avantages pour l'élaboration d'approches de recherche. La théorie des systèmes en constitue une base solide. Deuxièmement, le cadre peut être utilisé comme modèle pour structurer la charte de gouvernance des données d’une organisation. Certaines chartes sont disponibles sur des sites web, principalement dans le domaine de la santé. Toutefois, elles se concentrent sur les obligations légales et les questions de conformité. Fournir un tel cadre aux experts chargés de rédiger une charte facilite la recherche de l'exhaustivité. Nous avons vérifié sur certaines chartes existantes que notre cadre couvre tous les sujets qu'elles contiennent. Troisièmement, le cadre de référence fournit une base pour l'évaluation de la gouvernance des données. Dans cette optique, chaque dimension peut être associée à des tests d'audit et à des mesures. Nous avons commencé à définir les questions d'audit avec un panel de praticiens. À titre d'exemple, nous avons affiné les objectifs de valeur, de risque et de coût en élaborant des typologies pour chacun d'entre eux. La valeur englobe plusieurs composantes : commerciale, intrinsèque, de performance, économique, etc. Chacune d'entre elles peut être mesurée à l'aide d'un indicateur spécifique. Enfin, le cadre de référence peut servir de point de départ au développement d’un modèle de maturité de la gouvernance des données. De nombreux modèles de maturité sont proposés dans la littérature. Ils sont utilisés pour évaluer, par exemple, la manière dont les parties prenantes comprennent et adoptent les concepts et les méthodes. Pour développer un modèle de maturité, chaque dimension du cadre de référence doit être déclinée en niveaux de maturité et associée à un ensemble de mesures.

5. Conclusion et recherche future

Cette recherche a pour but de structurer le domaine de la gouvernance des données en développant un cadre conceptuel fondé sur la théorie des systèmes qui définit ses dimensions clés. Ce faisant, nous avons répondu au besoin d'une vision holistique qui puisse guider à la fois les chercheurs dans l'élaboration d'hypothèses et les praticiens dans l'organisation de leur gouvernance des données. Cette recherche a abouti à un cadre conceptuel pour la gouvernance des données mais également mis en évidence plusieurs axes dans lesquels des recherches supplémentaires sont nécessaires. Ainsi, une validation empirique du cadre conceptuel auprès de praticiens est en cours, ainsi que l'opérationnalisation d'un modèle de maturité de la gouvernance des données. L'applicabilité du cadre conceptuel à d'autres types de gouvernance (informatique, connaissance, entreprise, etc.) est une autre voie de recherche future.

Remerciements. *Les auteurs remercient les relecteurs pour leurs précieux conseils et les partenaires de la Chaire ESSEC Stratégie et gouvernance des données et de l'IA au sein de laquelle cette recherche a été menée.*

Bibliographie

Abraham, R., Schneider, J., & Vom Brocke, J. (2019). Data governance: A conceptual framework, structured review, and research agenda. *International journal of information management*, 49, 424-438.

Akoka, J., & Comyn-Wattiau, I. (2019, June). Évaluation de la gouvernance de l'information: une approche holistique. In AIM 2019: 24ème Conférence de l'Association Information et Management.

Alhassan, I., Sammon, D. and Daly, M. (2018), "Data governance activities: a comparison between scientific and practice-oriented literature", *Journal of Enterprise Information Management*, Vol. 31 No. 2, pp. 300-316. https://doi.org/10.1108/JEIM-01-2017-0007

Alhassan, I., Sammon, D., & Daly, M. (2016). Data governance activities: an analysis of the literature. *Journal of Decision Systems*, 25(sup1), 64-75.

Alhassan, I., Sammon, D., & Daly, M. (2019). Critical success factors for data governance: a theory building approach. *Information systems management*, 36(2), 98-110.

Al-Ruithe, M., Benkhelifa, E., & Hameed, K. (2019). A systematic literature review of data governance and cloud data governance. *Personal and ubiquitous computing*, 23, 839-859.

Bennett, S. (2017). What is information governance and how does it differ from data governance? *Governance Directions*, 69(8), 462-467.

Brennan, R., Quigley, S., De Leenheer, P., & Maldonado, A. (2018). Automatic extraction of data governance knowledge from slack chat channels. In OTM Confederated International Conferences" On the Move to Meaningful Internet Systems" (pp. 555-564). Cham: Springer International Publishing.

Davidson, E., Wessel, L., Winter, J. S., & Winter, S. (2023). Future directions for scholarship on data governance, digital innovation, and grand challenges. *Information and Organization*, 33(1), 100454.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS quarterly*, 75-105.

Janssen, M., Brous, P., Estevez, E., Barbosa, L. S., & Janowski, T. (2020). Data governance: Organizing data for trustworthy Artificial Intelligence. *Government information quarterly*, 37(3), 101493.

Khatri, V., & Brown, C. V. (2010). Designing data governance. *Communications of the ACM*, 53(1), 148-152.

Le Moigne, J. L. (1994). La théorie du système général : théorie de la modélisation. PUF.

Merkus, J., Helms, R., & Kusters, R. J. (2019, May). Data Governance and Information Governance: Set of Definitions in Relation to Data and Information as Part of DIKW. In ICEIS (2) (pp. 143-154).

Nguyen, T. C. (2016). Information governance and management in the context of Gov 2.0 (Doctoral dissertation, Swinburne).

Otto, B. (2011). Organizing data governance: Findings from the telecommunications industry and consequences for large service providers. *Communications of the Association for Information Systems*, 29(1), 3.

Paparova, D., Aanestad, M., Vassilakopoulou, P., & Bahus, M. K. (2023). Data governance spaces: the case of a national digital service for personal health data. *Information and Organization*, 33(1), 100451.

Rastogi, N., & Trivedi, M. K. (2016). PESTLE technique–a tool to identify external risks in construction projects. *International Research Journal of Engineering and Technology (IRJET)*, 3(1), 384-388.

Simon, H. A. (1996). The architecture of complexity: hierarchic systems.

Skyttner, L. (1996). *General systems theory: An introduction*. London: Macmillan Press.

Smallwood, R. F. (2019). *Information Governance: Concepts, Strategies and Best Practices*, Wiley.

Tallon, P. P., Ramirez, R. V., & Short, J. E. (2013). The information artifact in IT governance: Toward a theory of information governance. *Journal of management information systems*, 30(3), 141-178.

Thor, A., Bornmann, L., Haunschild, R., & Leydesdorff, L. (2021). Which are the influential publications in the Web of Science subject categories over a long period of time? CRExplorer software used for big-data analyses in bibliometrics. *Journal of Information Science*, 47(3), 419-428.

Van Eck, N. J., & Waltman, L. (2013). VOSviewer manual. Leiden: Universiteit Leiden, 1(1), 1-53.

Weber, K., Otto, B., & Österle, H. (2009). One size does not fit all---a contingency approach to data governance. *Journal of Data and Information Quality (JDIQ)*, 1(1), 1-27.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1), 1-9.