INFormatique
des ORganisations
et Systèmes
d'Information
et de Décision

iNforSiD

TOULOUSE

26 • 29 | MAI 2026

Présidente du Comité de Programme :
Marlène Villanova

Présidente du forum JCJC :
Manuele Kirsch Pinheiro

Présidents du Comité d'Organisation :
Franck Ravat et Jiefu Song

Actes de la 44ème édition de la conférence INFORSID

UNIVERSITÉ TOULOUSE CAPITOLE

IRIT

cnrs

cimi TOULOUSE

inforsid.fr

L’association INFORSID

Siège Social : 44, Chemin de la Caille - 31750 Escalquens
Web : https://inforsid.fr/

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

Composition du bureau

Présidente	Agnès FRONT - LIG, Université Grenoble Alpes
Vice-présidente	Cécile FAVRE - ERIC, Université Lumière Lyon 2
Trésorier	Sébastien LABORIE - LIUPPA, Univ. de Pau et des Pays de l’Adour
Secrétaire	Rebecca DENECKERE - CRI, Univ. Paris 1 Panthéon-Sorbonne
Chargé d'animation scientifique	Olivier TESTE - IRIT, Université de Toulouse
Chargé de communication	Cyril FAUCHER - L3i, La Rochelle Université

Présidents d’honneur

Jean-Bernard CRAMPES	(Toulouse)
Gilles ZURFLUH	(Toulouse)
André FLORY	(Lyon)
Claude CHRISMENT	(Toulouse)
Michel SCHNEIDER	(Clermont-Ferrand)
Corine CAUVET	(Aix-Marseille)
Chantal SOULE-DUPUY	(Toulouse)
Dominique RIEU	(Grenoble)
Régine LALEAU	(Paris)
Franck RAVAT	(Toulouse)

Préface

Depuis sa première édition en 1982, le congrès INFORSID réunit chaque année la communauté recherche en INFormatique des ORganisations et Systèmes d'Information et de Décision. Lieu d'échange privilégié pour aborder les défis, discuter les problématiques, débattre des opportunités ou encore partager des solutions, le congrès est accueilli cette année à Toulouse pour sa 44ème édition. Après Nice (2010), Paris (2013), Grenoble (2016), Nantes (2018), INFORSID se tiendra conjointement à la conférence internationale RCIS (IEEE Research Challenges in Information Science) dont c'est la 20ème édition.

L'organisation conjointe de RCIS et d'INFORSID se concrétise par plusieurs temps scientifiques communs. Ainsi, 3 conférences invitées sont au programme et je remercie sincèrement leurs orateurs pour nous faire l'honneur de leur présence :

- Cyril Labbé, Université Grenoble Alpes / LIG nous parlera d'« Information Systems for Scientific Publications: Detection of Research Rubbish and More »
- Barbara Pernici, Politecnico di Milano invite au questionnement suivant : « Are We Ready for Data-Centric AI? An Information Systems Engineering Perspective »
- Mario Lasso Cisneros, Airbus, Toulouse traitera de « Commercial Aircraft development: geometrical variation information management from architecture trough development, manufacturing and operations»

Deux sessions labellisées "communes RCIS / INFORSID" voient 4 articles présentés, retenus par les deux comités de programme. Un cinquième article proposé dans cette catégorie double soumission fait l'objet d'une présentation au forum RCIS. Les participants des deux conférences pourront donc se retrouver lors de ces différents temps et prolonger leurs échanges au moment des pauses, déjeuner et événements sociaux partagés tout au long des journées.

Cette année, l'appel à communication d'INFORSID a donné lieu à 53 soumissions d'articles couvrant un grand nombre des problématiques liées à l'ingénierie des systèmes d'information et aux nouveaux développements méthodologiques et techniques qui l'entoure. Plusieurs catégories de soumission étaient possibles qui ont donné lieu à :

- 14 propositions au titre d'une Double Soumission INFORSID/RCIS
- 19 soumissions d'articles Recherche Papier Long
- 3 soumissions d'articles Recherche Papier Long Court
- 13 soumissions d'articles Internationaux[1]
- 4 soumissions d'articles Démo

Le processus de soumission incluait la possibilité pour un article en Double soumission non retenu pour RCIS d'être considéré dans la catégorie Recherche Papier Long. Sur les 9 papiers concernés, les auteurs de 7 d'entre eux ont opté pour cette possibilité, montant à 25 le nombre d'articles de cette catégorie pour INFORSID.

Chacun des articles soumis a été évalué par trois membres du Comité de Programme. Sur la base des revues réalisées, le Conseil du Comité de Programme a procédé à la sélection de :

- 15 articles Recherche Papier Long
- 2 articles Recherche Papier Court
- 8 articles Internationaux
- 4 articles Démo

Ces travaux sont donc présentés lors du congrès et les textes associés peuvent être retrouvés dans ces actes, qui incluent également des résumés pour les 5 articles de la catégorie Double soumissions INFORSID/RCIS. Les versions intégrales de ces articles sont proposées dans les actes de RCIS.

Les actes de la conférence INFORSID témoignent cette année encore du panel riche et diversifié des thématiques abordés au sein de la communauté INFORSID. En plus des 2 sessions communes avec RCIS, 7 sessions thématiques sont proposées : Systèmes de recommandation, Ingénierie des SI, Représentation et Gestion de Connaissances, Enjeux de soutenabilité et de cybersécurité, SI et approches LLM, Explicabilité et Confiance, et Fouille de processus. Les auteurs de démos logicielles les introduiront en session plénière et les présenteront sur le stand dédié.

Comme de coutume, INFORSID 2026 débute également par une journée d'ateliers. Pour deux d'entre eux, il s'agit d'une première édition :

- WISE : 1st International Workshop on Information Systems : Social and Environmental Sustainability, conjointement supporté par RCIS et organisé par Claudia Roncancio et Monica Vitali.
- Enseigner la conception et le développement des systèmes d'information à l'ère des LLM : ingénierie des SI, des processus et des services, proposé par François Charoy et Chihab Hanachi,

[1] Articles publiés en 2025 dans d'excellentes conférences ou revues internationales.

et se tiendra la

- 3ème édition de l'Atelier Systèmes d'Information et Humanités Numériques (SI-HN), porté par Stéphane Lamassé et Cédric du Mouza.

Autre temps fort de la conférence, le forum Jeunes Chercheurs et Jeunes Chercheuses est au programme, orchestré par Manuele Kirsch Pinheiro, avec cette année la participation de 12 doctorants et doctorantes dont les travaux ont été retenus. Ils et elles se prêteront lors d'une session plénière dédiée à l'exercice de la présentation éclair et de la critique constructive entre pairs, sous le regard bienveillant de l'ensemble des participants.

Enfin, l'Assemblée Générale de l'Association INFORSID est l'occasion de remettre le Prix de Thèse 2026. Ce prix est attribué à Adrien Petralia pour sa thèse intitulée "Apprentissage profond pour la caractérisation des séries temporelles de consommation électrique". Soutenu le 7 mai 2025, ce travail a été mené à l'Université Paris Cité, en partenariat avec EDF R&D. Il fait l'objet d'une présentation du lauréat en séance plénière. Un résumé de la thèse apparait dans les actes.

A quelques jours d'ouvrir la conférence, je tiens à remercier tous ceux et celles, et ils sont nombreux, qui contribuent à cette 44ème édition, par leur investissement depuis plusieurs mois et par leur participation à venir à ces journées :
— les 141 auteurs pour avoir produit et soumis des articles de recherche, et ceux et celles qui parmi eux viendront les présenter
— les 42 membres du comité de programme pour le travail d'évaluation de grande qualité qu'ils ont mené
— les 7 membres du conseil du comité de programme pour leur action ayant mené à la constitution du programme
— les porteurs et porteuses d'ateliers pour leur rôle précieux d'animation de la communauté
— les conférenciers et conférencière invités pour leur disponibilité et le partage de leur regard expert
— les présidents et présidentes de session qui ont répondu très gentiment à ma sollicitation
— les membres du bureau de l'Association INFORSID pour leur confiance et leur accompagnement sans faille. Merci Agnès Front, Cécile Favre, Sébastien Laborie, Rebecca Deneckère, Olivier Teste et Cyril Faucher. A l'heure où le mandat de notre présidente se termine, je profite également de cette tribune pour la remercier plus particulièrement pour son action à la tête de l'association. Je souhaite bien sûr le meilleur au bureau dans la poursuite de son engagement.
— les membres du comité d'organisation pour leur travail, sous la direction de Franck Ravat et Jiefu Song. Merci en particulier pour votre efficacité et la fluidité de notre collaboration, la coordination avec les équipes RCIS et pour avoir cette année

encore conjugué avec brio recherche et convivialité, si caractéristique des éditions d'INFORSID.

Enfin, merci à vous tous, participants et participantes, de longue date ou nouveaux arrivants. Soyez tous et toutes les bienvenus à l'édition 2026 d'INFORSID. J'aurai beaucoup de plaisir à vous renouveler mes remerciements de vive voix sous ce beau soleil toulousain de mai !

Marlène Villanova

Présidente du comité de Programme
d'INFORSID 2026

Comités

Comité de programme

Bernd	AMANN	LIP6 - Sorbonne Université
Pierre-Emmanuel	ARDUIN	DRM - Université Paris-Dauphine
Sarah	BOURAGA	EM Normandie
Sandra	BRINGAY	LIRMM - Université Paul Valéry Montpellier
Jean-Michel	BRUEL	IRIT - Université de Toulouse 2
Guillaume	CABANAC	IRIT - Université de Toulouse
Sylvain	CASTAGNOS	LORIA - Université de Lorraine
Ronan	CHAMPAGNAT	L3i - La Rochelle Université
Anouck	CHAN	ONERA
Alexandre	CHANSON	LIFAT - Université de Tours
Max	CHEVALIER	IRIT - Université de Toulouse
Nadine	CULLOT	LIB - Université de Bourgogne
Rebecca	DENECKERE	CRI - Université Paris 1 Panthéon Sorbonne
Cécile	FAVRE	ERIC - Université Lyon 2
Agnès	FRONT	LIG - Université Grenoble Alpes
Alexis	GUYOT	LIS - Aix Marseille Université
Lilia	GZARA	DISP - INSA Lyon
Stéphane	JEAN	LIAS - Université de Poitiers
Gérald	KEMBELLEC	CNAM
Éric	KERGOSIEN	GERIICO - Université de Lille
Elena	KORNYSHOVA	CNAM
Sébastien	LABORIE	LIUPPA - Université de Pau et des Pays de l'Adour
Anne	LAURENT	LIRMM - Université de Montpellier
Sabine	LOUDCHER	ERIC - Université Lyon 2
Maude	MANOUVRIER	LAMSADE - Université Paris-Dauphine-PSL
Maxime	MASSON	LIUPPA - Université de Pau et des Pays de l'Adour
André	PENINOU	IRIT - Université Toulouse 2
Thomas	POLACSEK	ONERA
Christophe	PONSARD	CETIC - Université de Namur - Belgique
Jolita	RALYTÉ	ISS - Université de Genève
Claudia	RONCANCIO	LIG - Université Grenoble Alpes
Philippe	ROOSE	LIUPPA - Université de Pau et des Pays de l'Adour
Irina	RYCHKOVA	CRI - Université Paris 1 Panthéon Sorbonne
Ines	SAAD	ESC AMIENS
Camille	SALINESI	CRI - Université Paris 1 Panthéon-Sorbonne

Christian	SALLABERRY	LIUPPA - Université de Pau et des Pays de l'Adour
Marinette	SAVONNET	LIB - Université de Bourgogne
Sana	SELLAMI	LIS - Aix Marseille Université
Olivier	TESTE	IRIT - Université Toulouse 2
Robert	VISEUR	Université de Mons
Yves	WAUTELET	KU Leuven
Cecilia	ZANNI-MERK	LITIS - INSA Rouen Normandie

Conseil du Comité de Programme

Sylvain	CASTAGNOS	LORIA - Université de Lorraine
Cécile	FAVRE	LIRIS - Université Lyon 2
Maude	MANOUVRIER	LAMSADE - Université Paris-Dauphine-PSL
Éric	KERGOSIEN	GERIICO - Université de Lille
Sébastien	LABORIE	LIUPPA - Université de Pau et des Pays de l'Adour
Christophe	PONSARD	CETIC - Université de Namur - Belgique
Olivier	TESTE	IRIT Toulouse

Comité d'Organisation

Le congrès INFORSID 2026 a été organisé par l'Institut de Recherche en Informatique de Toulouse (IRIT) et l'Université Toulouse Capitole.

Co-présidents du comité d'organisation

Franck RAVAT
Jiefu SONG

Membres

Eric ANDONOFF	Moncef GAROUANI
Lydie BALLABRIGA	Chihab HANACHI
Guillaume CABANAC	Gilles HUBERT
Pierre-Paul CAVALLERA	Imen MEGDICHE
Anouck CHAN	Manon PREDHUMEAU
Yohann CHASSERAY	Ronan TOURNIER
Ophélie FRAISIER-VANNIER	Yanpei WANG

Conférences Invitées communes RCIS – INFORSID
- Résumés -

Cyril Labbé (Univ. Grenoble Alpes, LIG)

Information Systems for Scientific Publications: Detection of Research Rubbish and More

Scientific publications and the citations accompanying them were originally intended to disseminate knowledge. However, large and distributed information systems of various forms are now treating them as (ac)counting units for evaluation forming the basis of a wide range of metrics and rankings that shape the core logic of publishing activity. This has led to new types of bizarre artifacts that can be automatically detected: meaningless publications, tortured phrases, irrelevant or sneaked references, obvious errors, and more. Automatic analysis of scientific text, targeting specific misconducts, provides an actionable tool for detecting inappropriate and problematic publications.

> **Cyril Labbé** received a PhD in computer science (1999) from University of Grenoble. He is a tenured professor in computer science and co-PI of the ERC-Synergy *"NanoBubbles: How, how, when and why does science fail to correct itself?"*. His work on automatic detection of meaningless scientific papers, has led to retractions or withdrawals of countless computer science and bio-medical publications. He created the "scigen detection" and "seek&blastn" softwares, participated to the "Problematic Paper Screener" website, uncover the existence of "sneaked reference" and did create Ike Antkare, a fictitious scientist, that had once (dixit Google Scholar) an astonishing h-index.

Barbara Pernici (Politecnico di Milano)

Are We Ready for Data-Centric AI? An Information Systems Engineering Perspective

Drawing on research carried out in several interdisciplinary projects in domains including emergency-related social media analysis, chemical engineering, and justice, this keynote addresses the challenges of applying Information Systems Engineering approaches in heterogeneous contexts. Particular attention will be devoted to domain understanding, data collection, and data preparation, with a focus on the relationships between conceptual modeling, large language models, and data-

centric AI. The talk will also address the main open sustainability issues related to these activities.

> **Barbara Pernici** is Professor of Computer Engineering at Politecnico di Milano. Her research focuses on adaptive information systems design, data and information quality, energy efficiency in information systems, and social media analysis. She has led the Information Systems Group at Politecnico di Milano in several national and European projects, including Crowd4SDG, ECO2Clouds, GAMES, and Discount Quality for Responsible Data Science. She has served as elected Chair of IFIP TC8 Information Systems and of IFIP Working Group 8.1 on Information Systems Design.

Mario Lasso Cisneros (Airbus)

Commercial Aircraft development: geometrical variation information management from architecture trough development, manufacturing and operations

The nowadays large commercial aircraft industry requires the involvement of a large number of manufacturers worldwide. From early concepts stages through detailed design and manufacture of aircraft components, the geometrical variation management is strongly linked to the global industrial strategy. How to coordinate the different actors across the supply chain? How to find the balance to deliver against the current demand of new aircraft? What are the different alternatives to produce components and assemble them? The co-design of Aircraft and the industrial system that manufacture at the expected quality, rate and performance is the key. This keynote will focus on one of the aspects that critically enable these ambitions is how the geometrical variation is managed along the different manufacturing and assembly steps.

> **Mario Lasso Cisneros** holds a masters degree in advanced design techniques (2010) from the University of Bordeaux. He has worked within the Airbus' Design Office addressing the geometrical consistency of the different design principles of aircraft fuselage, wings and propulsion systems. Also, he has worked in the different Airbus Plants in France, Germany, Spain and the United Kingdom where he has addressed the challenge of aligning aircraft geometrical requirements to industrial system's manufacturing capabilities for the latest developments of Airbus Commercial aircrafts (A350, A320NEO, A330NEO). He is currently leading the Airbus Business Process architecture for Geometrical Variation Management in order to harmonise the way of working within the Airbus Commercial Aircraft Division.

Prix de thèse 2026

Actes de la conférence

Session 1 - Systèmes de recommandation

Session 2 - Commune RCIS

Session 3 - Ingénierie des SI

Session 4 - Représentation et Gestion de Connaissances

Session 7 - Commune RCIS

Session 8 - SI et approche LLM

Session 9 - Explicabilité et Confiance

Session 10 - Fouille de processus

Prix de thèse

VISA: Valoriser les Intérêts Spécifiques des Autistes pour accéder à des métiers qualifiés

Iakov Belkov[1], Maha Massoudi[1], Ali Hassan[1], Patrice Darmon[1], Patrick Marc Korenblit[2]

1. Research & Innovation - CGI
10-12 Cours Michelet, 92800 Puteaux
{iakov.belkov, maha.massoudi, a.hassan, patrice.darmon}@cgi.com

2. Association APTE autisme
2 Rue Wilfrid Laurier, 75014 Paris
patrick.korenblit@wanadoo.fr

RESUME. *VISA est un projet d'orientation professionnelle pour les personnes autistes. Il valorise les intérêts spécifiques et les compétences, tout en intégrant les sensibilités sensorielles (bruit, lumière, interactions), les préférences et les niveaux d'expertise. Deux moteurs de recommandation sont proposés : le premier à base de recherche multicritère avec une saisie guidée et le deuxième à base d'une recherche sémantique avec une saisie libre en langage naturel. Notre proposition vise à améliorer la pertinence des résultats de recommandation et à réduire la latence, le coût et l'empreinte carbone.*

ABSTRACT. *VISA is a career guidance project for autistic individuals. It emphasizes users' specific interests and competencies while accounting for sensory sensitivities (noise, lighting, interpersonal interaction), personal preferences, and levels of expertise. Two recommendation engines are provided: a multi-criteria retrieval approach with guided input and a semantic retrieval approach with free-form natural-language input. Our approach aims to improve recommendation relevance while reducing latency, cost, and carbon footprint.*

MOTS-CLÉS : *autisme ; recommandation de métiers ; intérêts spécifiques ; sensibilités sensorielles ; recherche sémantique ; recherche multicritère ; embeddings*

KEYWORDS: *autism ; job recommendation ; specific interests ; sensory sensitivities ; semantic search ; multi-criteria search ; embeddings*

1. Introduction

L'accès à l'emploi qualifié reste un enjeu majeur pour les personnes autistes. Au-delà des barrières systémiques (discrimination, inadéquation des environnements de travail, difficultés d'ajustement), l'orientation professionnelle souffre souvent d'un manque d'outils capables de relier des profils atypiques à des métiers concrets et réalistes. Un levier fréquemment observé dans l'accompagnement est la valorisation des *intérêts spécifiques* (IS) : ces centres d'intérêt, parfois très marqués, peuvent favoriser l'acquisition de compétences transférables (rigueur, expertise thématique, apprentissage autonome), mais ils sont rarement exploitables par les moteurs de recherche d'emploi classiques.

L'association APTE-Autisme souhaite mettre en place un site web capable de recommander des métiers qualifiés. Dans ce contexte, l'enjeu de VISA (**V**aloriser les **I**ntérêts **S**pécifiques des **A**utistes) est de proposer un moteur de recommandation capable de traiter des requêtes en langage naturel et de recommander des *métiers qualifiés*, en tenant compte à la fois des IS, des compétences et des *sensibilités sensorielles* (SS) des personnes autistes, susceptibles d'impacter la compatibilité avec certains métiers (bruit, lumière, interactions sociales, etc.). La problématique n'est pas uniquement le *matching* sémantique : il s'agit d'intégrer des critères hétérogènes et partiellement flous (préférences, expertise, contraintes sensorielles), tout en conservant une recommandation compréhensible et exploitable par des utilisateurs non experts (personnes autistes autonomes, parents, aidants, professionnels).

Nous adressons deux enjeux :

1) un **enjeu social**, en outillant les parents, les accompagnants et les professionnels pour orienter les personnes autistes vers des métiers cohérents avec leurs IS, aptitudes et caractéristiques psycho-sociologiques ;

2) un **enjeu économique**, en favorisant l'accès à l'emploi qualifié et la réduction du non-emploi (souvent estimé très élevé ; par exemple, on cite fréquemment l'ordre de grandeur de 400 000 adultes autistes en âge de travailler dont jusqu'à 95% seraient au chômage ou bénéficiaires d'allocations[1]).

Nous proposons deux approches complémentaires. **La première** repose sur une interaction *guidée* : l'utilisateur sélectionne ses IS, ses compétences et ses SS via des listes prédéfinies. Il peut paramétrer en détail ses niveaux d'expertise et ses degrés de préférence pour chaque compétence. Un moteur de recommandation multicritère prend ces paramètres en compte, calcule un score de pertinence et retourne un classement (top-k) de métiers, en pénalisant les incompatibilités liées aux SS. **La seconde approche** vise à lever les limites d'un référentiel des IS fermé. Elle repose sur une expression *libre* des IS et des compétences en langage naturel. Un moteur de recommandation interroge un référentiel des métiers vectorisé via une recherche sémantique, avant un reclassement tenant compte des SS.

1. https://informations.handicap.fr/a-autistes-chomage-docu-brise-cliches-13342.php

L'article est structuré de la façon suivante. La section 2 présente l'état de l'art des systèmes de recommandation de métiers. La section 3 décrit le modèle de données. Les sections 4 et 5 sont consacrées à nos moteurs de recommandation multicritères à base de préférences et sémantique respectivement. Nous présentons l'implémentation de nos travaux dans la section 6. Ensuite, nous détaillions nos expérimentations dans la section 7. Nous concluons dans la section 8.

2. État de l'art

Dans cette section nous présentons les différentes catégories de l'état de l'art des recommandations de métiers.

Approches classiques : Les plateformes de recherche d'emploi (par exemple France Travail, Indeed) reposent historiquement sur des requêtes syntaxiques (mots-clés, filtres simples). Cette logique souffre d'une limite structurelle : l'équivalence sémantique n'est pas garantie (synonymes, variations d'intitulés, granularité des compétences), ce qui peut dégrader le rappel et la précision. Un même métier peut se formuler via des descriptions variées, de plus des profils comparables (CV) s'expriment très différemment. Cela augmente l'impact de cette limite. Ce qui nécessite (i) des représentations sémantiques robustes, (ii) des mécanismes d'appariement expressifs, et (iii) des garanties d'explicabilité et d'équité (Barrak *et al.*, 2022; Khelkhal and Lanasri, 2025).

Approches avec clustering et classification : Une première famille d'approches vectorise les contenus puis structure l'espace via du clustering. Mhamdi *et al.* (2020) proposent un système de recommandation fondé sur le clustering d'offres d'emploi, construit à partir d'attributs extraits des descriptions, puis enrichi par des signaux comportementaux des utilisateurs (candidatures, likes, évaluations) afin de produire une liste de recommandations top-n.

La solution TeamBuilder de Darmon *et al.* (2018) se basent sur une ontologie des compétences IT pour faire l'analyse sémantique des CV et des offres d'emploi de l'entreprise afin de faire une recommandation bi-directionnelle. Cette proposition a été amélioré en intégrant un modèle de préférences pour définir les profils des candidats (Slama and Darmon, 2021). Sur la base de cette modélisation, un algorithme de correspondance et de notation à base de logique floue est ensuite appliqué pour sélectionner les k meilleurs résultats personnalisés.

Dans une logique centrée sur l'intitulé des métiers IT, Rahman *et al.* (2025) introduisent JoTPaRS, exploitant une classification multi-label et une décision inspirée de la théorie des jeux pour choisir un intitulé de métier optimal selon les compétences et l'expérience de l'utilisateur. Ces deux approches ont une contrainte forte : elles couvrent uniquement des métiers dans le domaine IT. Le recrutement étant un marché à deux côtés, Alsaif *et al.* (2022) proposent une recommandation bi-directionnelle et ils calculent une similarité vectorielle de compétences de candidats et des offres.

Approches fondées sur les transformers et les représentations sémantiques : Les modèles à base de transformers permettent d'aligner CVs et emplois dans un espace vectoriel commun avec une capacité supérieure à gérer paraphrases et variations lexicales que les approches basées sur le clustering et la classification. CareerBERT projette CV et métiers dans un espace d'embedding partagé en s'appuyant sur La taxonomie européenne des compétences, qualifications et professions (ESCO) pour standardiser les intitulés des métiers Rosenberger *et al.* (2025). L'utilisation de la taxonomie réduit l'ambiguïté et fournit un cadre stable pour la recommandation.

Au-delà de l'alignement embedding, certains travaux font un pré-traitement afin de synthétiser les CVs ou les offres. El-Deeb *et al.* (2025) montrent que la synthétisation automatique (BART/T5/Pegasus) peut condenser les offres et augmenter significativement la pertinence de la recommandation, mais cette synthétisation peut supprimer des contraintes critiques. Liu *et al.* (2025) proposent une compréhension de CV renforcée par LLM pour des systèmes à grande échelle, interprétant les intentions et les critères pour améliorer le classement.

Explicabilité, équité et adaptation inclusive : En recrutement, l'explicabilité est une condition d'acceptabilité (audit, contestation, confiance). Même dans des approches comme CareerBERT (Rosenberger *et al.*, 2025), l'interprétabilité est identifiée comme axe prioritaire. Barrak *et al.* (2022) souligne la nécessité de chaînes de décision justifiables garantissant la traçabilité des traitements et la compréhension des appariements (CV/offre). *Smart-Hiring* (Khelkhal and Lanasri, 2025) vise une plateforme explicable de bout en bout, couplant extraction d'information et justification des scores.

Les besoins ne sont pas uniquement algorithmiques : la recherche d'emploi mobilise des interactions avec l'entourage (parents, aidants, conseillers) et des contraintes de communication. Ragozin *et al.* (2024) analysent ces pratiques chez des personnes autistes et proposent un prototype pour mieux soutenir la collaboration (planification, communication, préparation par étapes, soutien mutuel au sein d'une communauté neurodiversifiée), afin d'améliorer l'expérience de recherche d'emploi collaborative des personnes autistes.

En résumé, d'une part, les travaux portant sur la recherche de métiers pour les personnes autistes se concentrent sur l'accompagnement, sans proposer de moteur de recommandation de métiers. D'autre part, les travaux proposant des moteurs de recommandation de métiers ne prennent pas en compte les caractéristiques spécifiques des personnes autistes, telles que les intérêts spécifiques (IS) et les sensibilités sensorielles (SS), susceptibles d'impacter négativement les recommandations.

Pour surmonter ces limites, nous proposons un nouveau moteur de recommandation, VISA, spécifiquement conçu pour les personnes autistes.

3. Modèle de données

Soient :

– Niv={Faible, Modérée, Forte} : l'ensemble de niveaux,

– Imp={Négociable, Non-négociable} : l'ensemble de niveaux d'importance,

– SS : l'ensemble de sensibilités sensorielles.

Définition 1. Un *Métier* M_j est défini par :

– Nom^{M_j} : le nom du métier.

– Desc^{M_j} : la description du métier.

– Comp^{M_j} $= \{Comp_1^{M_j}, Comp_2^{M_j}, ..., Comp_n^{M_j}\}$ est un ensemble de compétences requises pour pratiquer le métier.

– $Comp^{M_j} \rightarrow (Niv, Imp)$: une fonction qui associe à chaque compétence un niveau d'expertise et un niveau d'importance. Par exemple, pour le métier "Secrétaire", le niveau d'importance associé à la compétence "Excel" est "Non-négociable", tandis que le niveau d'expertise requis est "Faible". Autrement dit, il est très important (Non-négociable) pour ce métier d'avoir des connaissances de base (Faible) en Excel.

– SS$^{M_j} = \{SS_1^{M_j}, SS_2^{M_j}, ..., SS_q^{M_j}\}$ est un ensemble de sensibilités sensorielles qu'on devrait supporter pour pratiquer le métier.

– $SS^{M_j} \rightarrow (Niv, Imp)$: une fonction qui associe à chaque sensibilité sensorielle un niveau d'intensité et un niveau d'importance.

Définition 2. Un *Intérêt spécifique* IS_k chez une personne autiste est défini par :

– Nom^{IS_k} : le nom de l'intérêt spécifique (IS).

– Desc^{IS_k} : la description de l'IS.

– Comp^{IS_k} $= \{Comp_1^{IS_k}, Comp_2^{IS_k}, ..., Comp_r^{IS_k}\}$ est l'ensemble de compétences acquises grâce à l'IS.

– $Comp^{IS_k} \rightarrow (Niv, Imp)$: une fonction qui associe à chaque compétence un niveau d'expertise et un niveau de préférence de l'autiste. Par exemple, même si une personne autiste a un niveau d'expertise "Faible" en "Programmation", elle peut avoir un niveau de préférence très élevé (Non-négociable) pour trouver un métier qui nécessite de pratiquer cette compétence.

Définition 3. Une *Sensibilité sensorielle* SS_a chez un autiste est définie par une fonction : $SS \rightarrow (Niv)$ qui associe à chaque sensibilité sensorielle un niveau d'intensité (Faible, Modérée, Forte).

La Figure 1 présente notre modèle de données. Dans notre modèle, nous ne collectons aucune donnée personnelle sur les autistes, c'est pourquoi la classe "User (Autiste)" ne comprend aucun attribut. Un autiste peut avoir plusieurs IS et plusieurs SS à la fois. Dans les classes d'énumération "Niveau" et "Importance", nous avons associé à chaque libellé une valeur numérique. Ces valeurs représentent le poids de chaque libellé par rapport aux autres. Ainsi, un niveau "Fort" est deux fois plus important qu'un niveau "Modéré" et six fois plus qu'un niveau "Faible". De la même manière, une importance "Non-négociable" est cinq fois plus importante que "Négociable". Ces valeurs ont été validées par les experts métier.

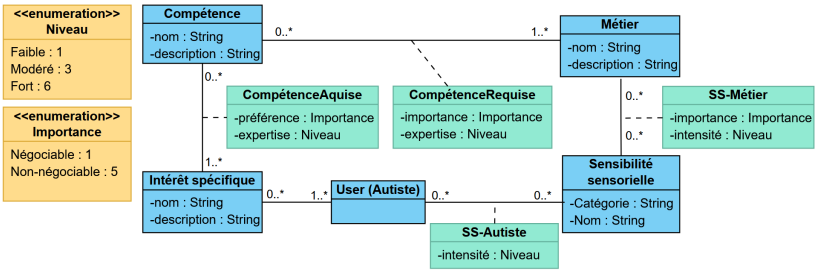


FIGURE 1. *Modèle des données*

4. Moteur de recherche multicritères à base de préférences

L'algorithme 1 détaille notre moteur de recherche multicritère. Il prend en entrée (I) la liste des compétences acquises grâce aux IS et (II) les SS de l'autiste. Et il donne en résultat une liste de métiers recommandés. Cet algorithme extrait d'abord la liste des métiers (possibles) dont la liste de compétences comprend une ou plusieurs des compétences acquises de l'autiste (ligne 1). Pour chacun de ces métiers (ligne 2), on initialise d'abord les variables utilisées pour calculer le score du métier (lignes 3-4). Puis, on parcourt les compétences (ligne 5) afin d'accumuler les niveaux d'expertise demandés (ligne 6) et les niveaux d'importance (ligne 7) des compétences. Ces deux cumuls seront utilisés pour calculer le score final du métier. Ensuite, lorsque l'autiste a la compétence considérée (ligne 8), on calcule :

– un score pour son niveau d'expertise (pondéré par l'importance de la compétence) par rapport au niveau demandé pour le métier (lignes 9-10),

– un score de sa préférence par rapport au niveau d'importance de cette compétence (lignes 11-12).

Pour chaque SS que l'on devrait supporter pour pratiquer le métier (ligne 13), si l'autiste y est sensible (ligne 14), alors on calcule un score de SS en se basant sur le niveau d'intensité de cette SS chez l'autiste et pour le métier à la fois. Ce score impacte négativement le score du métier (ligne 15). Ce score est alors intégré au cumul d'expertise de l'autiste via une contribution pondérée par l'intensité et l'importance (lignes 16-17).

Enfin, on calcule le score final du métier comme une combinaison des contributions normalisées de préférence et d'expertise (ligne 18). Si ce score dépasse (ou atteint) le seuil fixé, le métier est ajouté à la liste des recommandations avec son score (lignes 19-20). Cette liste est ordonnée selon un score décroissant (ligne 21).

Discussion : Contrairement à la recherche par mots-clés classique, qui reposent directement sur les termes saisis par l'utilisateur, nous identifions la liste des compétences acquises d'un autiste grâce au référentiel des IS. Par exemple, pour un IS "Sport",

nous pouvons identifier une longue liste de compétences : "préparation d'itinéraires", "navigation en montagne", "nutrition", "techniques de natation", "gestion du stress", "agilité physique", etc. Toutes ces compétences se trouvent dans le référentiel des métiers, associées aux différents métiers. Ces derniers sont alors identifiés comme métiers "possibles", même s'ils ne relèvent pas du domaine du sport. Ainsi, l'autiste peut choisir le niveau d'expertise et de préférence pour chaque compétence. Il peut même supprimer une compétence de sa liste s'il considère ne pas la posséder.

Algorithm 1: Recherche multicritères

Input: CompétencesAcquises, SS_Autiste
Output: MétiersRecommandés

PossibleMétiers ← Métiers.findByListCompName(CompétencesAcquises);
foreach *métier* ∈ *PossibleMétiers* **do**
  sumImportance ← 0 ; sumExpertise ← 0;
  userSumPréférence ← 0 ; userSumExpertise ← 0;
  foreach *Comp* ∈ *métier.CompétenceRequiseList* **do**
    sumExpertise← sumExpertise + (Comp.expertise×Comp.importance);
    sumImportance ← sumImportance + Comp.importance;
    if *Comp* ∈ *CompétencesAcquises* **then**
      scoreExpertise
      $\leftarrow \min\left(\frac{\text{CompétencesAcquises[Comp].expertise}}{\text{Comp.expertise}}, 1\right)$;
      userSumExpertise ← userSumExpertise + scoreExpertise × Comp.expertise × Comp.importance;
      scoreImportance
      $\leftarrow \min\left(\frac{\text{CompétencesAcquises[Comp].préférence}}{\text{Comp.importance}}, 1\right)$;
      userSumPréférence ← userSumPréférence + scoreImportance × Comp.importance;
  foreach *ss* ∈ *métier.SS-Métier* **do**
    if *ss* ∈ *SS_Autiste* **then**
      scoreSS $\leftarrow -1 \times \frac{\text{SS\_Autiste[ss].intensité + ss.intensité}}{2 \times \text{maxIntensité}}$;
      userSumExpertise ← userSumExpertise + scoreSS × ss.intensité × ss.importance;
      sumExpertise ← sumExpertise + (ss.intensité × ss.importance);
  score $\leftarrow \frac{1}{2}\left(\frac{\text{userSumPréférence}}{\text{sumImportance}} + \frac{\text{userSumExpertise}}{\text{sumExpertise}}\right)$;
  if *score* ≥ *scoreSeuil* **then**
    MétiersRecommandés.add(métier, score);
return *MétiersRecommandés.orderBy(score, desc)*

5. Moteur de recherche sémantique

Notre *moteur de recherche sémantique* est un système de recommandation qui s'appuie sur une représentation vectorielle sémantique des métiers (base de données vectorielle) afin de dépasser une recherche strictement lexicale par mots-clés. Ce moteur représente la requête utilisateur (IS et compétences en langage naturel) et les fiches métiers en vecteurs dans un même espace vectoriel, où la proximité géométrique reflète une proximité de sens. La recommandation est alors formulée comme un problème de recherche des plus proches voisins : pour une requête donnée, le système récupère les k métiers dont la similarité sémantique est maximale, puis applique, lorsque nécessaire, des traitements complémentaires, notamment un filtrage des résultats ambigus et un reclassement intégrant les sensibilités sensorielles.

Dans la suite de cette section, nous détaillons d'abord notre base de données vectorielle, puis nous proposons une méthode de recommandation préliminaire. Cette méthode fait l'objet d'une étude de validation basée sur un jeu de données représentatif.

5.1. *Base de données vectorielle des métiers*

Un *embedding* est une représentation vectorielle dense, apprise par un modèle neuronal, qui projette chaque texte dans un espace continu où la distance entre vecteurs reflète la similarité de sens. Lors de l'entraînement, le modèle est optimisé pour rapprocher les textes jugés similaires (p. ex. descriptions proches ou paraphrases) et éloigner les textes non liés, ce qui structure l'espace vectoriel. Les avancées en *embeddings* multilingues, comme Yu *et al.* (2025), permettent de les étendre sans compromis entre langues, facilitant le *matching* dans des contextes internationaux. Il est désormais possible de générer des *embeddings* robustes et sémantiquement précis pour des textes en français, permettant une indexation et un calcul de similarité directement dans cette langue, sans recourir à une traduction intermédiaire. Notre base vectorielle des métiers est construite en trois phases :

1) Une phase de préparation des données consiste à récupérer les fiches métiers et leurs métadonnées depuis la base applicative, puis à constituer une représentation textuelle exploitable pour l'indexation (découpage en *chunks*).

2) Une phase de vectorisation consiste à générer des *embeddings* à partir des fiches métiers au moyen d'un modèle d'*embedding* servi localement via Ollama, configuré ici avec un modèle de type *snowflake2*.

3) Une phase d'indexation consiste à construire une structure de recherche des k plus proches voisins à l'aide de FAISS Douze *et al.* (2025), destinée à accélérer la récupération des k métiers les plus similaires à une requête utilisateur.

5.2. *Méthode de recommandation*

La Figure 2 présente notre méthode de recommandation préliminaire. Cette méthode

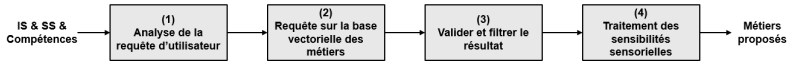


FIGURE 2. *Méthode de recommandation préliminaire*

constitue la pierre angulaire de notre méthodologie de recherche scientifique. Elle démarre à partir d'une requête utilisateur avec les IS et les compétences en langage naturel et une liste de SS. Elle contient quatre étapes :

1) La requête est analysée afin d'identifier et extraire les informations pertinentes et de produire une représentation exploitable.

2) Cette représentation alimente une interrogation de la base vectorielle des métiers, via un calcul de similarité pour récupérer les métiers les plus proches dans l'espace d'embedding.

3) Les métiers candidats retournés sont validés et filtrés selon des critères de qualité et de cohérence.

4) Un module dédié effectue le traitement des sensibilités sensorielles, en ajustant le score des métiers qui nécessitent l'exposition à des facteurs incompatibles avec le profil sensoriel.

Le processus produit enfin une liste ordonnée de métiers recommandés, correspondant au compromis optimal entre la pertinence sémantique de la requête et la compatibilité sensorielle. Notre méthode de recherche se base sur une étude d'évaluation fondée sur un jeu de données représentatif. Cette étude a été menée afin de valider la méthode de recommandation. Les critères de validation choisis sont la pertinence des résultats de recommandation, la latence, le coût et l'empreinte carbone.

5.2.1. *Jeu de données*

Afin de valider la robustesse du moteur de recherche face à des formulations hétérogènes, nous avons constitué un jeu de tests composé de 45 cas de recherche. Chaque cas est défini par une combinaison de trois composantes en langage naturel :

– **un intérêt spécifique** : nous considérons trois modalités d'expression :

1) des mots clés,

2) une formulation courte (une phrase),

3) une formulation moyenne (un paragraphe de deux à trois phrases).

– **des compétences liées à l'intérêt spécifique** : nous considérons quatre modalités d'expression :

1) absence de compétences,

2) mots-clés,

3) une formulation courte (une phrase),

4) une formulation moyenne (un paragraphe de deux à trois phrases).

– **des compétences non liées à l'intérêt spécifique** sont exprimées avec les mêmes quatre modalités que celles liées à l'intérêt spécifique.

Ce protocole permet de simuler des requêtes réalistes variant en longueur et en précision, tout en introduisant un facteur de bruit contrôlé via l'ajout de compétences non liées à l'intérêt spécifique. Au total, un cas peut ainsi générer jusqu'à 48 combinaisons de requêtes, correspondant au produit des modalités d'expression des trois composantes, soit 2,160 requêtes possibles sur l'ensemble des 45 cas.

Les requêtes générées sont ensuite vectorisées et utilisées pour interroger la base vectorielle des métiers. Dans la suite de la section, nous détaillons les expérimentations d'évaluation des quatre étapes de la méthode de recommandation préliminaire.

5.2.2. *Expérimentation sur l'analyse de la requête de recherche*

La fonctionnalité d'analyse de requête est assurée par un LLM. Elle a été utilisée afin de limiter l'impact du bruit et de la variabilité de la saisie en langage naturel sur la recherche sémantique. Les champs saisis par l'utilisateur, regroupant intérêts spécifiques et compétences, étaient soumis à un modèle de langage chargé de produire une reformulation courte et centrée sur les éléments pertinents. Cette reformulation est ensuite utilisée comme requête pour interroger la base vectorielle. Les expérimentations menées sur notre jeu de données n'ont pas montré d'amélioration significative de la pertinence et ont parfois dégradé les résultats en supprimant des éléments discriminants. En conséquence, cette fonctionnalité n'est pas activée dans la version finale, ce qui réduit la latence, le coût et l'empreinte carbone associés aux appels au LLM.

5.2.3. *Expérimentation sur les requêtes dans la base vectorielle des métiers*

Chaque fiche métier est transformée en un *embedding* de d dimensions, obtenu à l'aide d'un modèle d'*embeddings* servi localement via Ollama. Afin de réaliser une comparaison sémantique, les vecteurs métiers $v_i \in \mathbb{R}^d$ sont normalisés selon la norme L_2 lors de la construction du référentiel des métiers. De même, la requête utilisateur est vectorisée par le même modèle et normalisée, ce qui permet d'évaluer la proximité requête–métier via la similarité cosinus. Dans ce cadre, la similarité cosinus entre un vecteur requête q et un vecteur métier v_i est équivalente à leur produit scalaire :

$$\begin{aligned} \mathrm{sim}(q, v_i) = \cos(q, v_i) &= \frac{q \cdot v_i}{\|q\|_2 \|v_i\|_2} \\ &= q \cdot v_i \quad \text{si } \|q\|_2 = \|v_i\|_2 = 1. \end{aligned}$$

Le référentiel est indexé par FAISS via un index `IndexFlatIP`, qui effectue une recherche des k plus proches voisins selon le produit scalaire. Étant donné une requête q, FAISS retourne les k métiers maximisant $\mathrm{sim}(q, v_i)$, ainsi que les scores associés.

Un seuil minimal de similarité est alors appliqué afin d'éliminer les métiers dont la proximité sémantique avec la requête est insuffisante, garantissant que seuls des candidats présentant un niveau minimal de pertinence soient transmis aux étapes de filtrage suivantes. Les expérimentations menées sur notre jeu de données ont permis d'établir un seuil minimal de similarité à 0.4, ensuite confirmé par les experts métier. Les experts ont également fixé k à 20 métiers.

5.2.4. *Expérimentation sur la validation et le filtrage des résultats*

Ensuite, l'implémentation introduit une stratégie de validation. Les résultats de recherche par similarité sont soumis à une validation par un LLM de type GPT afin de réduire les faux positifs sémantiques. Les expérimentations menées sur le jeu de données ont montré que les métiers dont le score est élevé (supérieur à 0.7) passent systématiquement la validation. À l'inverse, les métiers dont le score est inférieur nécessitent cette étape de validation par un LLM, en raison d'un nombre important de faux positifs sémantiques. Ainsi, afin de réduire la latence, le coût et l'empreinte carbone, nous réservons les appels au LLM aux seuls cas ambigus (entre 0.4 et 0.7).

5.2.5. *Expérimentation sur le traitement des sensibilités sensorielles*

La prise en compte des sensibilités sensorielles est réalisée par un mécanisme de *re-scoring* après la validation et le filtrage. Les sensibilités de l'utilisateur sont prises en compte par le service ; chaque sensibilité est associée à une intensité. Pour chaque métier candidat, le moteur récupère les sensibilités associées au métier, puis calcule une pénalité multiplicative à partir d'une matrice croisant intensité et importance (Figure 3). Par exemple, si la personne autiste présente une sensibilité forte au bruit et que le métier implique une exposition forte non négociable, alors le score du métier est diminué de 35%. Les pénalités correspondant aux différentes sensibilités sont cumulées et plafonnées, puis appliquées au score de similarité afin de produire un score final. Les expérimentations menées sur le jeu de données ont permis aux experts métier de valider les pénalités proposées.

Cette méthode de recommandation permet de combiner une pertinence sémantique globale avec des contraintes de SS hétérogènes, sans imposer l'encodage direct des sensibilités dans l'espace vectoriel.

6. Implémentation

Les travaux sur l'implémentation peuvent être regroupés sous deux axes principaux (1) la collection des données et (2) le développement de l'application web.

6.1. *Collection des données*

Le projet VISA s'appuie sur deux référentiels de connaissances : un référentiel des métiers et un référentiel d'intérêts spécifiques (IS). Le référentiel des métiers a été

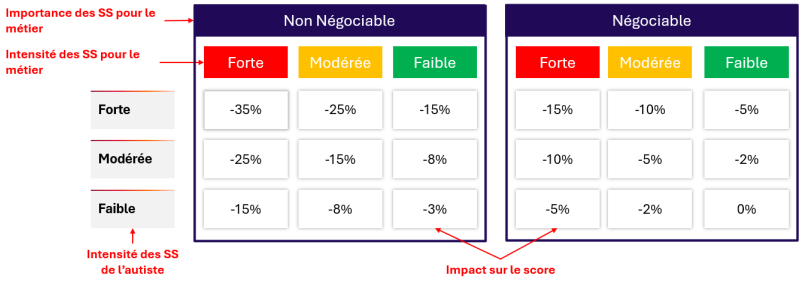


FIGURE 3. *Matrice de SS*

initialement constitué par extraction de fiches métiers issues de France Travail, puis enrichi pour répondre aux besoins des personnes autistes. Cet enrichissement consiste à associer à chaque compétence un niveau d'expertise attendu et un niveau d'importance (notamment pour distinguer les exigences centrales des compétences secondaires). En parallèle, nous avons intégré les sensibilités sensorielles (SS) comme des contraintes métier : pour chaque métier, certaines expositions (bruit, lumière, interactions, etc.) sont décrites par une intensité attendue et un niveau d'importance, afin de pouvoir pénaliser les recommandations incompatibles avec le profil sensoriel de l'utilisateur.

La préparation du référentiel des métiers inclut également un nettoyage des données destiné à améliorer la robustesse de la recommandation. Une attention particulière a été portée à la suppression de redondances et de doublons sémantiques dans les compétences entre métiers : des compétences formulées différemment mais équivalentes (variantes orthographiques, synonymie, granularité) peuvent artificiellement gonfler ou biaiser les scores. Le nettoyage vise donc à stabiliser le vocabulaire, à réduire le bruit et à limiter les effets de bord sur le classement.

Le référentiel des intérêts spécifiques (IS) répond à une logique différente : il ne s'agit pas seulement d'une liste de thèmes, mais d'un pont entre des centres d'intérêt et des compétences potentiellement transférables vers des métiers qualifiés. La définition des IS et l'établissement de liens plausibles avec des compétences et des métiers ont été menés dans une démarche semi-automatique : un LLM a été utilisé pour proposer des regroupements, des reformulations et des associations, puis ces propositions ont été revues et validées par des dizaines de volontaires ainsi que par des experts métier, afin de limiter les biais d'interprétation.

6.2. *Application web*

Notre application web est accessible sur le lien (`https://emploi-autisme.com`). L'architecture applicative de notre application web (Figure 4) repose sur une séparation explicite entre l'IHM, le backend métier, et le service IA, afin d'isoler la

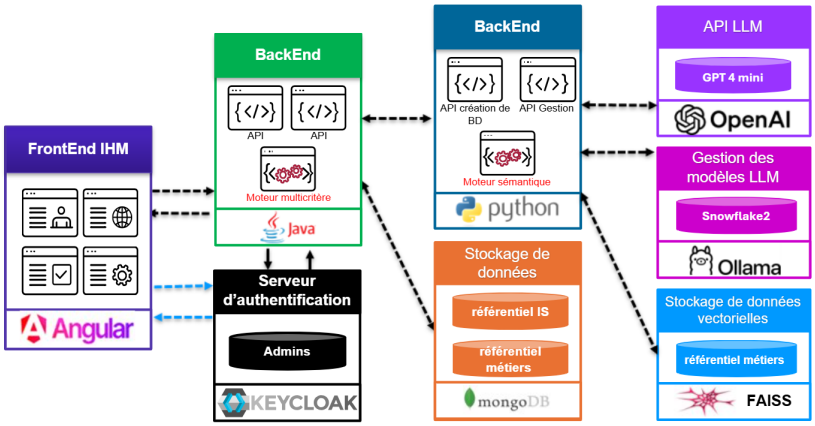


FIGURE 4. *Architecture de l'application web*

logique de recommandation des composants applicatifs classiques (gestion, persistance, administration). L'IHM est une application web (Angular) utilisée par deux profils : d'une part un administrateur (APTE-Autisme) qui gère les contenus et la qualité des référentiels, et d'autre part une personne autiste qui interroge le système. Côté utilisateur final, l'usage est conçu pour être anonyme, sans authentification, afin d'éviter la collecte de données personnelles.

Le backend applicatif (Java) porte les fonctions de gestion des données et expose les capacités du moteur multicritère, qui exploite les référentiels métiers & IS. Ces derniers sont persistées dans une base applicative (MongoDB). L'accès aux fonctions d'administration est protégé par un serveur d'authentification (Keycloak) tandis que la consultation et la recherche restent accessibles aux internautes (autistes).

Le moteur sémantique est implémenté comme un service IA dédié (Python), indépendant du backend métier. Il interagit avec un stockage vectoriel (FAISS) contenant la base vectorielle des métiers. Lorsque nécessaire, et selon les règles de gouvernance définies (coût/latence). Ce service peut aussi solliciter un LLM (via API OpenAI) pour lever des ambiguïtés sémantiques sur des métiers a recommander ayant un score pas assez élevé, tandis que la génération d'embeddings est assurée localement via un gestionnaire de modèles (Ollama) afin de réduire le coût, l'empreinte carbone et la dépendance à des services externes.

7. Expérimentations

Nous avons conduit une campagne de tests sur l'application VISA afin d'évaluer la qualité de la compatibilité des métiers proposés avec les données utilisateur (intérêts

spécifiques et compétences) et la corrélation entre le nombre des métiers recommandés et le nombre de mots saisis.

7.1. *Test qualité*

Protocole : Une première série d'essais a consisté à évaluer par les experts métier la compatibilité des métiers recommandés avec les entrées utilisateur (IS et compétences), selon quatre niveaux : *forte*, *modérée*, *faible* et *KO*. Pour cela, 16 cas de test ont été réalisés sur des domaines différents afin d'obtenir un panel de métiers. Les IS et les compétences ont été saisis sous forme de phrase(s). Les expertes métier ont réalisé un classement pour chaque métier selon les quatre niveaux.

Résultats : Selon le tableau 1, sur 114 métiers proposés issus de 16 tests, la compatibilité apparaît globalement solide. 83 recommandations sont classées *fortement compatibles* (72.8%) et 21 *modérées* (18.5%), soit 91.3% de résultats jugés pertinents. Les cas *faibles* représentent 7 recommandations (6.1%), tandis que les résultats *KO* restent marginaux avec 3 cas (2.6%). Au total, la proportion de recommandations non satisfaisantes (*faible* + *KO*) est limitée à 8.7%, ce qui confirme que le système produit majoritairement des recommandations cohérentes, avec une marge d'amélioration concentrée sur la réduction des cas *faibles* et *KO*.

Tests	**métiers proposés**	**Forte**	**Modérée**	**Faible**	**KO**
16	114	83	21	7	3
		72.8%	18.5%	6.1%	2.6%

TABLEAU 1. *Résultats du test qualité.*

7.2. *Test de corrélation*

Protocole : Les tests ont été réalisés en faisant varier la forme des requêtes utilisateur. Plusieurs cas couvrant des domaines différents ont été définis, avec saisie d'intérêts spécifiques et de compétences variés. Les requêtes ont été formulées sous différentes formes lexicales et syntaxiques (mots-clés, phrases courtes, formulations plus longues), tout en conservant le même profil, afin de mesurer l'effet de la rédaction sur les résultats retournés. Pour chaque requête, les métiers proposés et leurs scores ont été relevés.

Résultats : La figure 5 montre que l'augmentation du nombre de mots saisis tend à augmenter le nombre de métiers proposés et les scores associés, mais aussi la proportion de résultats moins pertinents. La qualité dépend fortement du choix des termes employés : la précision du vocabulaire a un impact plus déterminant que la taille du texte. Des formulations imprécises ou ambiguës dégradent sensiblement la pertinence des recommandations.

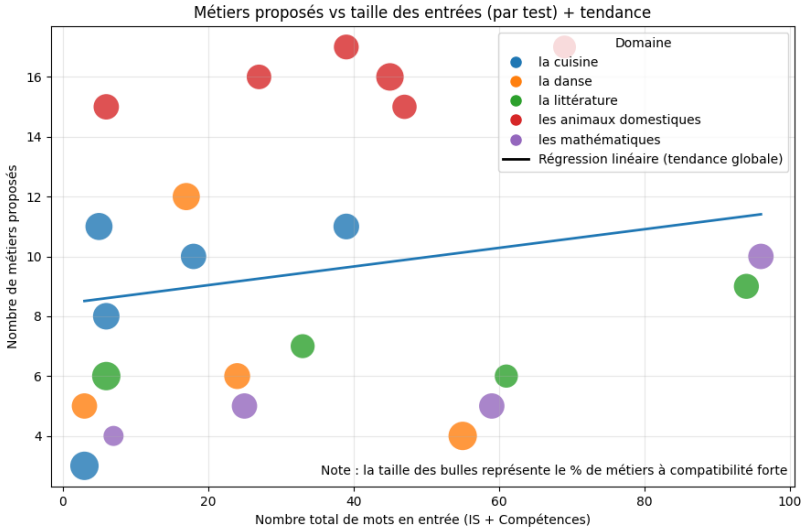


FIGURE 5. *Résultats du test de corrélation*

8. Conclusion

VISA est un dispositif d'orientation professionnelle dédié aux personnes autistes, qui vise à recommander des métiers qualifiés en valorisant les intérêts spécifiques et les compétences, tout en prenant en compte des contraintes de sensibilités sensorielles (bruit, lumière, interactions). Notre contribution repose sur deux moteurs complémentaires : un moteur multicritère à saisie guidée permettant de paramétrer niveaux d'expertise, préférences et incompatibilités, et un moteur de recherche sémantique à saisie libre fondé sur des *embeddings* et une indexation FAISS, complété par un mécanisme de validation par LLM réservé aux cas ambigus afin de limiter la latence, le coût et l'empreinte carbone. L'évaluation conduite sur un jeu de requêtes volontairement hétérogènes montre que cette combinaison améliore la robustesse face à la variabilité linguistique tout en conservant une capacité de filtrage et de re-classement adaptée aux contraintes sensorielles.

Les perspectives immédiates portent sur l'extension et la consolidation des référentiels (métiers, IS, SS), une évaluation en situation avec les utilisateurs finaux et les accompagnants, et le renforcement de l'explicabilité et de l'analyse des biais, afin d'accroître l'acceptabilité et la fiabilité des recommandations dans un cadre opérationnel.

Remerciement : Nous adressons nos remerciements à *PlanetHoster* pour la mise à disposition des trois machines virtuelles (VM) ayant permis la réalisation de ce travail. Nous exprimons également notre gratitude à *Malakoff Humanis* et à *l'association A4*

pour leur soutien financier. Nous remercions aussi *France Travail*, *KOREIS* et *Avencod* pour les ressources apportées dans le cadre du projet. Enfin, nous remercions les 33 experts et consultants qui ont participé à la réalisation de l'application web et à la collection des données.

Bibliographie

Alsaif S. A., Hidri M. S., Ferjani I., Eleraky H. A., Hidri A., « NLP-Based Bi-Directional Recommendation System : Towards Recommending Jobs to Job Seekers and Resumes to Recruiters », *Big Data and Cognitive Computing*, vol. 6, nº 4, p. 147, 2022.

Barrak A., Adams B., Zouaq A., « Toward a traceable, explainable, and fairJD/Resume recommendation system », *ArXiv*, 2022.

Darmon P., Mazouzi R., Manad O., Bentounsi M., « TeamBuilder : D'un moteur de recommandation de CV notés et ordonnés à l'analyse sémantique du patrimoine informationnel d'une société », *BDA 2018 – 34ème Conférence sur la Gestion de Données – Principes, Technologies et Applications*, Bucarest, Roumanie, October, 2018. hal-01980539.

Douze M., Guzhva A., Deng C., Johnson J., Szilvasy G., Mazaré P.-E., Lomeli M., Hosseini L., Jégou H., « The FAISS library », *IEEE Transactions on Big Data*, vol. , p. 1-17, 2025.

El-Deeb R. H., Abdelmoez W., El-Bendary N., « Enhancing E-Recruitment Recommendations Through Text Summarization Techniques », *Information*, 2025.

Khelkhal K., Lanasri D., « Smart-Hiring : An Explainable end-to-end Pipeline for CV Information Extraction and Job Matching », *Preprint, DOI : 10.48550/arXiv.2511.02537*, 2025.

Liu P., Shen J., Shen Q., Yao C., Kao K., Xu D., Arora R., Zheng B., Johnson C., Hong L., Wu J., Zhang W., « Powering Job Search at Scale : LLM-Enhanced Query Understanding in Job Matching Systems », *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, Association for Computing Machinery, New York, NY, USA, p. 4971–4975, 2025.

Mhamdi D., Moulouki R., Ghoumari M. Y. E., Azzouazi M., Moussaid L., « Job Recommendation based on Job Profile Clustering and Job Seeker Behavior », *Procedia Computer Science*, vol. 175, p. 695-699, 2020.

Ragozin K., Rough D., van Berkel N., Hettiachchi D., Kostakos V., « Collaborative Job Seeking for People with Autism : Challenges and Design Opportunities », *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, p. 1-18, 2024.

Rahman S., Nur F. N., Afridi A. S., Islam A. H. M. S., Sultana S., Moon N. N., « JoTPaRS : A Job Title Prediction and Recommendation System for IT Professionals », *Advances in Artificial Intelligence and Machine Learning*, vol. 5, nº 3, p. 4356-4378, 2025.

Rosenberger J., Wolfrum L., Weinzierl S., Kraus M., Zschech P., « CareerBERT : Matching resumes to ESCO jobs in a shared embedding space for generic job recommendations », *Expert Systems with Applications*, vol. 275, p. 127043, 2025.

Slama O., Darmon P., « A Novel Personalized Preference-based Approach for Job/Candidate Recommendation », *Research Challenges in Information Science*, p. 418-434, 2021.

Yu P., Merrick L., Nuti G., Campos D. F., « Arctic-Embed 2.0 : Multilingual Retrieval Without Compromise », *Second Conference on Language Modeling*, 2025.

Scores de substituabilité et de complémentarité entre items pour l'évaluation des systèmes de recommandation

Marie Griffon[1,2], Nicolas Delestre[1], Maxime Gueriau[1], Cecilia Zanni-Merk[1]

1. INSA Rouen Normandie, Univ Rouen Normandie, Université Le Havre Normandie, Normandie Univ, LITIS UR 4108
F-76000 Rouen, France
prenom.nom@insa-rouen.fr

2. TraceParts
Parc Éco Normandie
F-76430 Saint-Romain-de-Colbosc, France
mgriffon@traceparts.com

RESUME. *Les systèmes de recommandation (SRs) sont devenus des composantes essentielles des plateformes numériques. À mesure que leur impact économique s'accroît, l'évaluation de leurs performances ne peut plus se limiter aux seules métriques de précision. En particulier, les évaluations hors ligne existantes prennent peu en compte la notion d'utilité des recommandations du point de vue de l'utilisateur. Dans cet article, nous soulignons l'intérêt d'une évaluation orientée vers l'utilité pour les SRs et nous proposons deux scores permettant de quantifier les relations de substituabilité et de complémentarité entre items à partir de données de sessions d'utilisateurs, avec pour objectif de servir de fondement à la construction de nouvelles métriques d'évaluation hors ligne pour les SRs. Ces métriques ont pour objectif d'évaluer l'utilité des recommandations pour les utilisateurs ainsi que leur capacité à être considérées et utilisées comme un lot. Le recours à des scores de substituabilité et de complémentarité est motivé par l'observation que les interactions observées au sein d'une session traduisent un processus de décision orienté par les objectifs de l'utilisateur, conduisant à l'exploration d'items substituables et complémentaires. La formulation des scores proposés repose sur les co-occurrences d'items et sur la notion de contexte de session. Des expérimentations menées sur plusieurs jeux de données publics montrent que notre approche permet de capturer des relations de substituabilité et de complémentarité significatives entre items.*

ABSTRACT. *Recommender systems (RSs) have become essential components of online platforms. As their economic impact continues to grow, evaluating their performance can no longer be limited to accuracy-based metrics alone. In particular, existing offline evaluations largely overlook the usefulness of recommendations from the user's perspective. In this paper, we emphasize the importance of usefulness-oriented evaluation for RSs and propose two scores to quantify substitutability and complementarity relationships between items based on user session data, with the objective of serving as a foundation for the construction of new offline evaluation metrics for RSs. These metrics would be designed to assess both the usefulness of recommendations for users and their ability to be perceived and used as a coherent set. The use of substitutability and complementarity scores is motivated by the observation that interactions within a session reflect a decision-making process driven by the user's objectives, leading to the exploration of substitutable and complementary items. The proposed score formulation relies on item co-occurrences and on the notion of session context. Experiments conducted on several public datasets show that our approach is able to capture meaningful substitutability and complementarity relationships between items.*

MOTS-CLÉS : *Système de recommandation, Évaluation hors ligne, Relations d'item*

KEYWORDS: *Recommender system, Offline evaluation, Item relationships*

1. Introduction

Les systèmes de recommandation (SRs) jouent un rôle crucial dans la performance commerciale des plateformes en ligne modernes. Ils sont largement utilisés dans le e-commerce, le streaming de contenus multimédias, la publicité en ligne et les services numériques afin d'aider les utilisateurs à naviguer dans de vastes catalogues de contenus, de produits ou de services, désignés de manière générique comme des *items*. Ces systèmes visent également à stimuler à la fois l'engagement des utilisateurs et les ventes.

Des plateformes majeures telles qu'Amazon et Netflix s'appuient fortement sur des recommandations personnalisées, avec une part substantielle des interactions des utilisateurs et des revenus étant directement attribuée aux algorithmes de recommandation (MacKenzie *et al.*, 2013; Chen and Liu, 2017). En conséquence, les SRs sont passés du statut d'outils auxiliaires à celui de composants centraux des stratégies commerciales en ligne (Choi and Park, 2025).

À mesure que les SRs influencent de plus en plus les ventes, la consommation et l'expérience utilisateur, l'évaluation de leurs performances est devenue un enjeu central tant pour la recherche que pour l'industrie. Pour l'évaluation des SRs, trois cadres expérimentaux sont généralement distingués : les évaluations hors ligne, les études utilisateurs et les évaluations en ligne. Les évaluations hors ligne reposent sur des jeux de données préalablement collectés contenant des retours explicites (*e.g.* des notes attribuées aux *items*) ou implicites (*e.g.* les *items* achetés, consultés, ou consom-

més). Les études utilisateurs impliquent des expériences contrôlées menées auprès d'un nombre limité de participants. Elles nécessitent généralement un recrutement coûteux et une préparation importante, et peuvent introduire des biais comportementaux, les utilisateurs ayant tendance à modifier leur comportement lorsqu'ils savent qu'ils participent à une étude, un phénomène connu sous le nom d'« effet Hawthorne » (Landsberger, 1958). Les évaluations en ligne consistent à déployer le SR dans un environnement réel (Zangerle and Bauer, 2022). Bien que les évaluations en ligne offrent une estimation réaliste du comportement des utilisateurs, elles comportent des risques importants. En particulier, si un SR recommande un trop grand nombre d'*items* non pertinents ou incohérents lors d'une évaluation en ligne, la confiance des utilisateurs peut rapidement se détériorer, les amenant à ignorer les recommandations, même après une amélioration ou un redéploiement du système. Cette situation est largement considérée comme la plus critique et la moins acceptable dans les contextes commerciaux (Chen and Liu, 2017). Il est donc essentiel de mener des évaluations hors ligne capables d'évaluer à la fois la précision et la qualité des recommandations afin de limiter les risques liés à l'exposition des utilisateurs à des *items* potentiellement non pertinents.

Une telle démarche implique d'évaluer un SR à l'aide de plusieurs métriques complémentaires et de trouver un bon équilibre entre celles-ci, car une évaluation fondée sur une seule métrique est généralement biaisée et insuffisante pour refléter ses performances globales (Zangerle and Bauer, 2022). Or, la majorité des travaux existants dans la littérature des SRs s'appuient encore principalement sur des métriques axées sur la précision dans les évaluations hors ligne. Ces métriques évaluent la capacité d'un SR à prédire les interactions passées des utilisateurs avec les *items*. Ces évaluations reposent notamment sur des métriques de précision de prédiction, comme l'erreur absolue moyenne (MAE) ou l'erreur quadratique moyenne (RMSE), des métriques de prédiction de pertinence, telles que la Précision@k et le Rappel@k, ainsi que des métriques de classement, comme le gain cumulatif actualisé normalisé (NDCG) ou le rang réciproque moyen (MRR) (Zangerle and Bauer, 2022).

Bien que la précision ait longtemps été considérée comme l'objectif principal des SRs, plusieurs travaux soutiennent qu'elle ne suffit pas à évaluer la qualité des recommandations (McNee *et al.*, 2006; Ge *et al.*, 2010; Kaminskas and Bridge, 2016). En particulier, ces travaux mettent en avant la nécessité de prendre en compte des facteurs supplémentaires influençant la qualité des recommandations et d'évaluer les listes de recommandations dans leur ensemble, plutôt que les *items* individuellement. En réponse, de nouveaux objectifs, chacun capturant un aspect différent du comportement de recommandation, ont été introduits : la diversité, la sérendipité, la nouveauté et la couverture (Kaminskas and Bridge, 2016; Ge *et al.*, 2010; Zangerle and Bauer, 2022).

Malheureusement, les métriques couramment utilisées dans les évaluations hors ligne ne capturent pas explicitement deux aspects essentiels : (1) le fait que les recommandations répondent aux besoins actuels de l'utilisateur ; et (2) le fait que les recommandations constituent un ensemble d'*items* cohérents qui peut être considéré comme un lot par l'utilisateur.

Dans une démarche visant à combler ces lacunes, nous introduisons deux scores permettant d'identifier et de quantifier les relations de substituabilité et de complémentarité entre *items* à partir des interactions observées dans les sessions des utilisateurs. Ces scores constituent le socle de notre contribution et sont conçus pour servir de base à la construction de nouvelles métriques d'évaluation hors ligne pour les SRs. Le recours à des scores de substituabilité et de complémentarité est motivé par l'observation que les interactions observées au sein d'une session traduisent un processus de décision orienté par les objectifs de l'utilisateur, conduisant à l'exploration d'items substituables et complémentaires.

Dans la suite de cet article, la section 2 présente les travaux connexes relatifs à l'évaluation des SRs et à l'extraction des relations de substituabilité et de complémentarité entre *items*. La section 3 décrit notre approche. La section 4 présente les résultats expérimentaux. Enfin, la section 5 discute les résultats obtenus et expose les perspectives de recherche futures.

2. Travaux connexes

Il existe une vaste littérature consacrée à l'évaluation des SRs qui définit et catégorise les métriques d'évaluation. Certaines revues proposent une synthèse générale des métriques couramment utilisées et les regroupent en grandes catégories, par exemple les métriques de précision de prédiction, les métriques de classement, les métriques orientées *business*, etc (Zangerle and Bauer, 2022; Chen and Liu, 2017), tandis que d'autres se concentrent sur des sous-ensembles spécifiques de métriques (Ge *et al.*, 2010; Kaminskas and Bridge, 2016; Kunaver and Pozrl, 2017).

Parmi les métriques d'évaluation hors ligne recensées dans ces travaux, aucune ne permet d'évaluer explicitement si les recommandations sont utiles pour l'utilisateur, ni si les recommandations peuvent être perçues et utilisées comme un lot par l'utilisateur. Il existe des métriques axées sur la précision qui évaluent la pertinence d'*items* pour lesquels un retour explicite ou implicite de l'utilisateur est disponible, mais elles ne fournissent aucune information sur la pertinence des autres *items* présents dans la liste, pour lesquels aucun retour n'est observé. Dans ce contexte, nous cherchons à construire de nouvelles métriques d'évaluation hors ligne pour les SRs répondant à ces enjeux et fondées sur les relations entre les *items*.

Plusieurs travaux se sont intéressés à l'étude des relations de substituabilité et de complémentarité entre les *items* dans les SRs, motivés par leur importance en économie, en marketing et en nutrition.

Tian *et al.* (2021) proposent une approche fondée sur les graphes pour extraire des relations de substituabilité et de complémentarité entre produits à partir de données de transactions. Les données sont modélisées sous la forme d'un graphe biparti reliant les produits et les transactions, et les relations entre produits sont analysées à partir de leurs motifs de co-occurrence. Les produits complémentaires sont identifiés comme des *items* co-apparaissant significativement plus souvent que prévu dans les mêmes transactions,

tandis que les produits substituables sont définis comme des *items* co-apparaissant rarement mais partageant des schémas d'achat similaires. Bien que nous partagions ces hypothèses pour caractériser les relations entre *items* (*cf.* section 3.3), cette approche repose sur une modélisation sous forme de graphe biparti, ce qui en complexifie la mise en œuvre et en limite l'applicabilité à des cas d'usage réels à grande échelle.

Sun *et al.* (2017) proposent d'utiliser la co-occurrence d'*items* (*Item Co-occurrence*, IC) comme mesure pour quantifier les relations entre *items*. L'IC permet d'identifier des relations de complémentarité ou de substituabilité à partir des probabilités de co-occurrence des paires d'*items*. Cette mesure est ensuite exploitée dans un cadre de recommandation fondé sur les caractéristiques des *items* qui peuvent être organisées sous forme de hiérarchie (*e.g.* des vêtements pour femme divisés en catégories de style vestimentaire, puis divisés en sous-catégories de type de vêtements), afin d'améliorer à la fois la précision et l'interprétabilité des recommandations. Si l'IC constitue une mesure simple et efficace pour extraire des relations entre *items*, elle ne prend pas en compte le contexte dans lequel les interactions ont lieu.

Enfin, Akkoyunlu *et al.* (2017) étudient la substituabilité des *items* dans le domaine spécifique de la recommandation alimentaire. Les auteurs introduisent le concept de contexte alimentaire, défini comme l'ensemble des aliments consommés conjointement avec un aliment donné, ainsi que le concept de contexte de prise alimentaire, qui décrit les conditions de consommation telles que le type de repas ou les participants. Ils supposent que deux aliments sont substituables s'ils sont consommés dans des contextes alimentaires similaires, et que le degré de substituabilité dépend du contexte de prise alimentaire. Afin de limiter la complexité combinatoire du problème, l'approche repose sur une modélisation des repas sous forme de graphes et sur l'analyse de repas différant par un seul aliment. Cette stratégie, bien qu'efficace, complexifie l'implémentation de l'approche et conduit à ignorer certains contextes alimentaires potentiels d'un *item*. De plus, ce travail se limite à l'étude de la substituabilité et ne propose pas de score de complémentarité.

3. Approche

Dans cette section, nous introduisons les notations, le cadre formel et le problème de recommandation étudié. Ces éléments constituent le socle nécessaire à la définition ultérieure des scores.

3.1. *Notations et formulation du problème*

Soient I l'ensemble des *items* et S l'ensemble des sessions. Nous définissons une session comme un ensemble d'*items* avec lesquels l'utilisateur a interagi, délimité par des bornes définissant le début et la fin d'un événement donné (*e.g.* une transaction ou une consommation) (Wang *et al.*, 2021). La tâche de recommandation étudiée est la suivante : étant donnée une session $Q \in S$, on génère un ensemble d'*items* $R \subseteq I$,

appelé recommandations, susceptibles d'intéresser l'utilisateur. Notre approche est guidée par deux objectifs. D'une part, nous cherchons à évaluer l'utilité des recommandations R relativement à la session Q, c'est-à-dire à déterminer dans quelle mesure les recommandations répondent aux besoins de l'utilisateur. D'autre part, nous visons à évaluer la cohérence interne de R, en analysant si les recommandations forment un ensemble cohérent d'*items* liés par des relations significatives.

Dans une session, les utilisateurs ont tendance à explorer des *items* fortement liés, partageant une intention commune (Wang *et al.*, 2019). Ces interactions traduisent un processus de décision orienté par un ou plusieurs objectifs ponctuels. Dans ce contexte, une recommandation est considérée comme utile si elle entretient une relation significative avec au moins un des *items* déjà présents dans la session, en cohérence avec l'intention qui la sous-tend.

En économie, les relations entre deux *items* sont classiquement caractérisées par les notions de substituabilité et de complémentarité (Sun *et al.*, 2017; Nicholson and Snyder, 2012). Ces relations ne sont pas binaires : il existe différents degrés de substituabilité et de complémentarité selon les usages et les contextes. Considérons par exemple les habitudes de consommation du *beurre de cacahuète* et de la *confiture* au petit-déjeuner. Aux États-Unis, ces deux produits sont fréquemment consommés ensemble sur du *pain*, ce qui traduit une forte relation de complémentarité. En revanche, en France, le *beurre de cacahuète* et la *confiture* sont plus souvent consommés de manière alternative sur le *pain*, ce qui correspond à une relation de substituabilité. Cet exemple met en évidence que les relations entre deux *items* peuvent être complémentaires dans certains contextes et substituables dans d'autres.

Ainsi, notre approche vise à quantifier le degré de substituabilité et de complémentarité entre deux *items*. Étant donnée une base de données de sessions, nous cherchons à identifier ces relations à partir des interactions observées. Pour ce faire, nous nous appuyons exclusivement sur des informations contextuelles issues des sessions, à l'instar des travaux de Akkoyunlu *et al.* (2017), qui proposent un score de substituabilité fondé sur les usages et les contextes d'utilisation des *items*, sans recourir à leurs attributs intrinsèques.

3.2. *Définir le contexte*

Nous adaptons les deux notions de contexte introduites par Akkoyunlu *et al.* (2017) à notre cadre applicatif. Nous définissons ainsi deux contextes distincts : le *contexte d'un item dans une session* et le *contexte de la session*.

Le contexte d'un *item* dans une session correspond à l'ensemble des autres *items* avec lesquels l'utilisateur a interagi au cours de cette même session. Par exemple, en modélisant le petit-déjeuner *{café, pain, beurre, confiture}* comme une session, le contexte de l'*item café* dans cette session est *{pain, beurre, confiture}*. Plus formellement, pour une session $s \in S$ et un *item* $i \in s$, le contexte de l'*item* i dans la session s est défini par $c_i(s) = s \backslash \{i\}$, avec $c_i(s) \subset I$.

Le contexte de la session permet de caractériser le cadre dans lequel elle s'est produite. Il peut inclure, par exemple, des caractéristiques liées à l'utilisateur (âge, localisation), ainsi que les objectifs poursuivis pendant la session (prendre le petit-déjeuner, s'habiller, décorer, *etc.*) (Wang *et al.*, 2019; Adomavicius and Tuzhilin, 2010).

Le contexte d'un *item* dans une session et le contexte de la session sont tous deux essentiels à l'étude de la substituabilité des *items*. En effet, un *item* peut remplacer un autre uniquement si, d'une part, les deux *items* entretiennent des interactions similaires avec les autres *items*, c'est-à-dire s'ils partagent des *items* complémentaires communs (Tian *et al.*, 2021), et si, d'autre part, ils sont associés aux mêmes facteurs contextuels (Akkoyunlu *et al.*, 2017).

3.3. *Calcul d'un score de substituabilité et de complémentarité*

Deux *items* sont dits substituables lorsqu'ils remplissent une même fonction et peuvent être utilisés *l'un à la place de l'autre*, tandis que deux *items* sont complémentaires lorsqu'ils sont utilisés *conjointement* afin d'enrichir une même expérience.

Nos hypothèses, inspirées de Tian *et al.* (2021) et de Akkoyunlu *et al.* (2017), reposent sur l'idée que les relations de substituabilité et de complémentarité entre *items* peuvent être inférées à partir des interactions observées dans les sessions. Intuitivement, la substituabilité est associée à des interactions effectuées dans les mêmes contextes, tandis que la complémentarité se manifeste par des co-interactions fréquentes au sein d'une même session. Cependant, la co-interaction entre deux *items* tend à réduire leur substituabilité potentielle, car ces *items* peuvent être complémentaires. Inversement, le fait que deux items partagent les mêmes contextes dans les sessions tend à réduire leur degré de complémentarité et suggère davantage une relation de substitution. Considérons, à titre d'exemple, cinq petits-déjeuners : *{pain, beurre, café}*, *{pain, beurre, thé}*, *{pain, pâte chocolatée à tartiner, café}*, *{pain, beurre, jus de fruits}* et *{pain, pâte chocolatée à tartiner, thé}*. La co-interaction entre le *beurre* et le *pain* est observée dans 3 petits-déjeuners. Cette récurrence suggère une relation de complémentarité entre ces deux *items*, le *beurre* venant enrichir la consommation du *pain* dans le cadre du petit-déjeuner. Le *beurre* et la *pâte chocolatée à tartiner* apparaissent dans deux mêmes contextes, *{pain, café}* et *{pain, thé}*, tout en co-apparaissant rarement dans un même petit-déjeuner. Cette similarité de contexte, combinée à une faible co-interaction, suggère une relation de substituabilité : ces deux *items* remplissent un rôle comparable et sont utilisés comme des alternatives l'un à l'autre.

Nous proposons donc des scores capables de quantifier le degré de substituabilité ou de complémentarité entre deux *items*, tout en tenant compte de l'influence opposée de l'autre relation. Dans ce cadre, nous cherchons à définir des scores de substituabilité et de complémentarité satisfaisant les hypothèses suivantes :

(1) Deux *items* sont fortement substituables lorsque leurs contextes dans les sessions sont identiques ;

(2) Deux *items* sont moins substituables lorsque les utilisateurs interagissent fréquemment avec les deux au sein des mêmes sessions ;

(3) La substituabilité est une relation symétrique ;

(4) Deux *items* sont fortement complémentaires lorsque les utilisateurs interagissent fréquemment avec les deux au sein des mêmes sessions ;

(5) Deux *items* sont moins complémentaires lorsque leurs contextes dans les sessions sont identiques ;

(6) La complémentarité est une relation symétrique ;

(7) Les relations de substituabilité et de complémentarité varient selon le contexte des sessions.

Tout d'abord, nous définissons l'ensemble C_i comme l'ensemble des contextes de l'*item* i dans les sessions, c'est-à-dire :

$$C_i = \{c_i(s) \mid s \in S, i \in s\} \quad [1]$$

Pour deux *items* $q, r \in I$, les hypothèses (1) et (5) sont capturées par l'intersection de C_q et C_r. Si $|C_q \cap C_r|$ est élevé, alors les utilisateurs interagissent fréquemment avec q et r dans les mêmes contextes.

Ensuite, nous notons $A_{q:r}$ l'ensemble des contextes de q dans lesquels r apparaît :

$$A_{q:r} = \{c \in C_q \mid r \in c\} \quad [2]$$

La valeur $|A_{q:r}|$ indique le nombre de contextes de q dans les sessions qui contiennent r.

En tenant compte de ces considérations et en nous fondant sur Akkoyunlu *et al.* (2017), nous définissons le score de substituabilité entre deux *items* q et r ainsi :

$$subst(q, r) = \frac{|C_q \cap C_r|}{|C_q \cup C_r| + |A_{q:r}| + |A_{r:q}|} \quad [3]$$

Le score de substituabilité est égal à 1 lorsque les contextes de q et de r dans les sessions sont identiques et que $A_{q:r} = A_{r:q} = \emptyset$. Si les utilisateurs n'interagissent jamais avec q et r dans les mêmes contextes, alors le score est égal à 0. Plus la valeur $|A_{q:r}| + |A_{r:q}|$ est élevée, plus le score diminue, ce qui reflète l'hypothèse (2).

Pour le score de complémentarité, nous proposons d'abord un score d'association estimant la probabilité que l'*item* r apparaisse dans un contexte de l'*item* q, avec une pénalisation de l'intersection de C_q et C_r :

$$assoc(q, r) = \frac{|A_{q:r}|}{|C_q| + |C_q \cap C_r|} \quad [4]$$

L'association, telle que définie ici, est une relation asymétrique. Or, nous cherchons à calculer un score de complémentarité qui quantifie le degré d'association mutuelle

entre deux *items*. Nous combinons donc les deux scores d'association asymétriques en un score de complémentarité symétrique, qui ne prend une valeur élevée que lorsque les deux scores asymétriques sont élevés. Dans la littérature, la moyenne harmonique est une mesure de tendance centrale couramment utilisée pour combiner des taux, et elle est bien adaptée à l'agrégation de scores (Ravana and Moffat, 2009). Elle est également plus sensible aux déséquilibres importants entre les valeurs agrégées, ce qui signifie qu'une forte association asymétrique entre deux *items* fera diminuer le score.

Nous définissons ainsi le score de complémentarité comme la moyenne harmonique des deux scores d'association :

$$compl(q,r) = \frac{2 \times assoc(q,r) \times assoc(r,q)}{assoc(q,r) + assoc(r,q)} \quad [5]$$

Le score de complémentarité de q et r est égal à 1 lorsque r apparaît dans tous les contextes de q et que q apparaît dans tous les contextes de r, et lorsque $C_q \cap C_r = \emptyset$. Si un utilisateur n'interagit pas avec r et q dans une même session, alors le score est égal à 0. Plus $|C_q \cap C_r|$ est élevé, plus le degré de substituabilité de q et de r est élevé, et plus le score de complémentarité diminue, ce qui reflète l'hypothèse (5).

4. Expérimentations

Cette section présente deux expérimentations menées afin d'évaluer les scores de substituabilité et de complémentarité proposés. La première expérimentation porte sur le jeu de données INCA2 et vise à comparer notre score de substituabilité à des approches issues de la littérature. La seconde expérimentation s'appuie sur le jeu de données BundleRec et a pour objectif d'évaluer la capacité de nos scores à identifier des relations de substituabilité et de complémentarité validées par des annotations humaines.

4.1. *Expérimentation sur le jeu de données INCA2*

Dans cette première expérimentation, nous comparons notre score de substituabilité à deux approches de la littérature : le score de substituabilité proposé par Akkoyunlu *et al.* (2017), et l'*Item Co-occurrence* (IC) introduite par Sun *et al.* (2017).

Le jeu de données français INCA2[1] (Étude Individuelle Nationale des Consommations Alimentaires 2) est issu d'une enquête menée entre 2006 et 2007 sur les habitudes de consommation alimentaire en France métropolitaine (AFSSA, 2009). Des carnets alimentaires couvrant une période de sept jours ont été collectés auprès de 2 624 adultes et 1 455 enfants, sur plusieurs mois.

1. https://www.data.gouv.fr/datasets/donnees-de-consommations-et-habitudes-alimentaires-de-letude-inca-2-3

Chaque journée est structurée autour de trois repas principaux : le petit-déjeuner, le déjeuner et le dîner. Pour chacun de ces repas, le lieu de consommation (domicile, travail, école ou extérieur) ainsi que la compagnie (famille, amis, collègues ou seul) sont renseignés.

Les 1 343 aliments recensés sont organisés selon une hiérarchie comprenant 44 groupes et 111 sous-groupes. Dans cette expérimentation, nous retenons les sous-groupes, ou les groupes lorsque l'aliment n'appartient à aucun sous-groupe, afin de capturer des relations de substitution à la fois intra-groupes et inter-groupes. Le groupe des aliments non codifiés est exclu.

L'ensemble des repas est regroupé dans une base de données, que nous partitionnons selon des informations contextuelles. Seules les données concernant les adultes sont conservées. La base est divisée en trois jeux de données en fonction du type de repas : (i) petit-déjeuner et déjeuner, (ii) petit-déjeuner, et (iii) déjeuner.

4.1.1. *Méthodologie*

Nous comparons trois scores de substituabilité : (1) le score de substituabilité proposé par Akkoyunlu *et al.* (2017), (2) l'*Item Co-occurrence* (IC) introduit par Sun *et al.* (2017), et (3) notre score de substituabilité.

Contrairement à l'approche (1), nous ne mettons en place aucune stratégie de réduction de la complexité combinatoire des contextes ; l'ensemble des contextes des items dans les sessions est pris en compte dans notre approche.

L'IC (2) repose sur la probabilité conjointe que deux *items* soient dans une même session et prend des valeurs dans $[0, +\infty[$. Une valeur d'IC égale à 1 indique une relation d'indépendance. Une valeur inférieure à 1 traduit une relation de substituabilité, tandis qu'une valeur supérieure à 1 indique une relation de complémentarité.

4.1.2. *Résultats*

Le tableau 1 présente le classement des trois principaux substituts pour plusieurs aliments, obtenu à l'aide du score de substituabilité de Akkoyunlu *et al.* (2017), de l'IC de Sun *et al.* (2017) et de notre score, à partir du jeu de données regroupant les sessions de petit-déjeuner et de déjeuner. Le tableau 2 se concentre sur notre score et propose le même classement, calculé séparément sur des sessions de petit-déjeuner uniquement et de déjeuner uniquement, afin d'analyser l'influence du type de repas sur les relations de substituabilité. Dans le tableau 2, nous n'incluons pas les classements obtenus à l'aide des deux autres scores, car les tendances observées sont comparables à celles mises en évidence dans le tableau 1. Les couleurs présentes dans ces tableaux correspondent chacune à un exemple cité dans cette section.

Les résultats du tableau 1 montrent que les relations de substituabilité identifiées à l'aide de notre score sont globalement similaires à celles obtenues avec l'approche de Akkoyunlu *et al.* (2017)

	Substituts (classés par score)		
Aliment	Score de Akkoyunlu *et al.* (2017) ↑	IC (Sun *et al.*, 2017) ↓	Notre score ↑
Pain	Biscottes (0.2234) Viennoiserie (0.1359) Gâteaux (0.0745)	Hamburgers et hot-dog (0.0114) Sandwichs baguette (0.0922) Céréales chocolatées (0.0970)	Biscottes (0.0249) Viennoiserie (0.0173) Jus de fruits (0.0070)
Café	Thé (0.2799) Cacao (0.1729) Chicorée (0.1486)	Chicorée (0.0775) Thé (0.1085) Cacao (0.1870)	Thé (0.0388) Chicorée (0.0193) Cacao (0.0183)
Thé	Café (0.2799) Cacao (0.1721) Chicorée (0.1289)	Chicorée (0.0217) Bière (0.0445) Hamburgers et hot-dog (0.0566)	Cacao (0.0503) Café (0.0388) Chicorée (0.0370)
Cacao	Chicorée (0.2171) Café (0.1729) Thé (0.1289)	Huile (0.0043) Sauces (0.0054) Légumes fruits (0.0085)	Chicorée (0.0709) Thé (0.0503) Céréales sucrées (0.0346)
Beurre	Margarine (0.2413) Miel/confiture (0.0924) Pâte chocolatée à tartiner (0.0786)	Sandwichs baguette (0.0642) Barres chocolatées (0.1244) Céréales chocolatées (0.2196)	Margarine (0.0231) Miel/confiture (0.0172) Jus de fruits (0.0127)
Lait	Jus de fruits (0.1409) Yaourts (0.1264) Sucre (0.1089)	Pizzas (0.0175) Cocktails (0.0182) Fruits au sirop (0.0279)	Jus de fruits (0.0505) Viennoiserie (0.0392) Miel/confiture (0.0359)
Vin	Sodas (0.0814) Bière (0.0704) Eau du robinet (0.0412)	Céréales sucrées (0.0178) Pâte chocolatée à tartiner (0.0284) Lait (0.0479)	Sodas (0.0064) Bière (0.0042) Eau du robinet (0.0041)
Pizzas	Sandwichs baguette (0.2429) Autres sandwichs (0.1729) Plats à base de pâtes ou de pommes de terre (0.1513)	Lait (0.0175) Viennoiserie (0.0453) Aliment particulier (0.0532)	Sandwichs baguette (0.0929) Autres sandwichs (0.0533) Hamburgers et hot-dog (0.0372)
Pommes de terre	Pâtes (0.1111) Haricots verts et petits pois (0.0922) Riz (0.0602)	Céréales chocolatées (0.0379) Céréales aux fruits frais ou secs (0.0398) Pâte chocolatée à tartiner (0.0440)	Pâtes (0.0091) Haricots verts et petits pois (0.0068) Mélanges de légumes (0.0045)

TABLEAU 1. *Classement des trois principaux substituts selon le score de substituabilité de Akkoyunlu et al. (2017), l'IC (Sun et al., 2017) et notre score, pour plusieurs aliments, sur les sessions de petit-déjeuner et de déjeuner*

De plus, le fait de considérer l'ensemble des contextes des *items* dans les sessions permet d'identifier des relations supplémentaires. Par exemple, dans le tableau 2, des substituts du *vin* ou de la *pizza* sont identifiés pour le petit-déjeuner, alors qu'aucun n'est identifié à l'aide du score de substituabilité de Akkoyunlu *et al.* (2017).

Cependant, l'absence de mécanisme de filtrage ou de pondération des contextes des *items* dans les sessions peut également introduire des relations moins pertinentes. Par exemple, dans le tableau 1, des relations de substituabilité telles que *jus de fruits* avec *pain* ou *beurre* apparaissent pour le petit-déjeuner et le déjeuner, bien qu'elles reflètent moins fidèlement les habitudes alimentaires.

Dans le tableau 1, les relations de substituabilité obtenues avec l'IC reflètent encore moins les habitudes alimentaires. Par exemple, pour le *beurre*, les substituts identifiés à l'aide de l'IC incluent *sandwichs baguette*, *barres chocolatées* et *céréales chocolatées*, ce qui est peu cohérent d'un point de vue nutritionnel. À l'inverse, notre score permet d'identifier à la fois des substitutions inter-groupes, comme $\{pommes\ de\ terre \Rightarrow haricots\ verts\ et\ petits\ pois\}$, et intra-groupes, comme $\{pain \Rightarrow biscottes\}$.

Les substituts identifiés pour les boissons sont également cohérents, puisqu'ils correspondent majoritairement à d'autres boissons. Par exemple, pour le *vin*, les substituts obtenus sont *sodas*, *bière* et *eau du robinet* (*cf.* tableau 1). Bien qu'aucune information sémantique explicite sur les modes de consommation ne soit disponible dans le jeu de données, la prise en compte des contextes des aliments dans les sessions permet d'en récupérer une approximation. Ainsi, certaines substitutions concernent des aliments

	Substituts (classés par notre score ↑)	
Aliment	Petit-déjeuner	Déjeuner
Pain	Biscottes (0.1159) Viennoiserie (0.0731) Gâteaux (0.0314)	Yaourts (0.0031) Pommes de terre (0.0029) Fruits (0.0028)
Café	Thé (0.1245) Chicorée (0.0695) Cacao (0.0644)	Yaourts (0.0045) Sodas (0.0033) Fruits (0.0030)
Thé	Café (0.1245) Cacao (0.0686) Chicorée (0.0523)	Sodas (0.0049) Autres sandwichs (0.0048) Bière (0.0046)
Cacao	Chicorée (0.0811) Thé (0.0686) Café (0.0644)	Légumes secs (0.0019) / /
Beurre	Margarine (0.1392) Pâte chocolatée à tartiner (0.0502) Fruits (0.0454)	Yaourts (0.0033) Margarine (0.0031) Fruits (0.0028)
Lait	Yaourts (0.0684) Eau du robinet (0.0659) Jus de fruits (0.0620)	Beignets, crêpes et gaufres (0.0057) Autres sandwichs (0.0050) Autre lait (0.0045)
Vin	Condiments (0.0286) Margarine (0.0036) Eau de source (0.0031)	Sodas (0.0063) Eau du robinet (0.0044) Bière (0.0042)
Pizzas	Lait aromatisé (0.2000) Autres produits de panification (0.1667) Biscuits apéritif (0.1429)	Sandwichs baguette (0.0938) Autres sandwichs (0.0511) Hamburgers et hot-dog (0.0372)
Pommes de terre	/ / /	Pâtes (0.0091) Haricots verts et petits pois (0.0068) Mélanges de légumes (0.0045)

TABLEAU 2. *Classement des trois principaux substituts selon notre score, pour plusieurs aliments, sur les sessions de petit-déjeuner uniquement, et sur les sessions de déjeuner uniquement*

appartenant aux mêmes groupes nutritionnels, comme *pommes de terre* et *pâtes*, qui sont tous deux riches en amidon (*cf.* tableau 2).

Enfin, les résultats du tableau 2 montrent que le découpage du jeu de données selon la variable contextuelle du type de repas conduit à des relations de substituabilité différentes. Par exemple, le *thé* est substitué par du *café*, du *cacao* et de la *chicorée* au petit-déjeuner, tandis qu'au déjeuner il est substitué par des *sodas*, d'*autres sandwichs* ou de la *bière*. Ces observations confirment que les relations de substituabilité dépendent fortement du contexte des sessions (*cf.* hypothèse (7) à la section 3.3).

4.2. *Expérimentation sur le jeu de données BundleRec*

L'objectif de cette seconde expérimentation est d'identifier des relations de substituabilité et de complémentarité à partir de sessions d'utilisateurs issues de trois domaines : l'alimentation, l'électronique et l'habillement. Les relations identifiées sont comparées à une vérité terrain construite à partir des annotations humaines proposées dans l'article de Sun *et al.* (2024), consacré à la recommandation de *bundles*. Un *bundle* correspond à un ensemble d'*items* proposés ou vendus ensemble afin de répondre à un objectif spécifique (*e.g.* s'habiller pour un événement ou compléter son équipement

de *gaming*). La recommandation de *bundle* consiste à proposer automatiquement à un utilisateur un *bundle* adapté à son intention d'achat.

Le jeu de données BundleRec[2] est issu de l'article de Sun *et al.* (2024). Les auteurs ont mis en place une tâche de *crowdsourcing* visant à annoter des *bundles* potentiels ainsi que les objectifs associés, dissimulés dans les sessions d'utilisateurs.

Les sessions proviennent des jeux de données Amazon (He and McAuley, 2016) et couvrent trois domaines : l'électronique, l'habillement et l'alimentation. Les statistiques descriptives de ces jeux de données sont présentées dans le tableau 3.

	Électronique	Habillement	Alimentation
#Items	4 943	6 439	5 072
#Sessions	1 145	1 181	1 161
#Bundles	1 750	1 910	1 784
Taille moyenne des bundles	3,52	3,31	3,58

TABLEAU 3. *Statistiques du jeu de données BundleRec*

Dans cette expérimentation, nous considérons les *items*, et non pas les catégories d'*items*, et nous exploitons les sessions des utilisateurs issues des trois jeux de données afin d'identifier des relations de substituabilité et de complémentarité entre les *items* dans chacun des domaines. La vérité terrain est constituée de 8 554 paires uniques d'*items* pour l'habillement, 9 387 pour l'électronique et 9 462 pour l'alimentation, correspondant aux paires uniques présentes dans les *bundles* annotés.

4.2.1. *Méthodologie*

Nous comparons deux approches : (1) l'*Item Co-occurrence* (IC) de Sun *et al.* (2017) et (2) nos scores de substituabilité et de complémentarité.

La précision et le rappel sont utilisés comme métriques d'évaluation pour mesurer, pour chaque domaine, la proportion de paires correctement prédites :

$$Precision = \frac{\#Paires\ correct.\ pred.}{\#Paires\ pred.} \qquad Rappel = \frac{\#Paires\ correct.\ pred.}{\#Paires\ VT}$$

où $\#Paires\ VT$ est le nombre de paires de la vérité terrain, $\#Paires\ pred.$ le nombre de paires prédites, et $\#Paires\ correct.\ pred.$ le nombre de paires à l'intersection entre les paires prédites et celles de la vérité terrain.

Une précision élevée indique que peu de relations non annotées (ou non avérées) sont identifiées, tandis qu'un rappel élevé indique que la majorité des relations annotées sont retrouvées.

2. https://github.com/BundleRec/bundle_recommendation

4.2.2. *Résultats*

Les résultats obtenus pour les trois domaines sont présentés dans le tableau 4. Les deux approches identifient les mêmes relations de complémentarité, ce qui s'explique par le fait que notre score de complémentarité et l'IC reposent tous deux sur la co-occurrence d'items dans les sessions. En revanche, notre approche identifie en outre des relations de substituabilité : 6 dans le domaine de l'électronique, 5 dans l'alimentation et 2 dans l'habillement. À l'inverse, l'IC n'identifie que des relations de complémentarité, toutes associées à un score supérieur à 1.

Domaine	Méthode	#Paires VT	#Paires préd.	#Paires correct. préd.	Précision	Rappel
Électronique	IC (Sun *et al.*, 2017)	9 387	18 777	9 387	49.99%	100%
	Nos scores		18 783	9 387	49.98%	100%
Habillement	IC (Sun *et al.*, 2017)	8 554	19 697	8 554	43.43%	100%
	Nos scores		19 699	8 554	43.42%	100%
Alimentation	IC (Sun *et al.*, 2017)	9 462	18 029	9 462	52.48%	100%
	Nos scores		18 033	9 462	52.47%	100%

TABLEAU 4. *Précision et rappel obtenus par IC (Sun et al., 2017) et par nos scores dans chacun des domaines*

Le rappel maximal obtenu par les deux approches indique qu'elles identifient l'ensemble des paires de la vérité terrain. Cependant, les valeurs de précision montrent qu'environ la moitié des paires prédites ne figurent pas dans les annotations ou ne sont pas considérées comme avérées. Ces résultats montrent que notre approche est comparable à l'IC en termes de précision et de rappel, tout en présentant l'avantage supplémentaire d'identifier des relations de substituabilité entre items. Toutefois, la question de la pertinence des paires non correctement prédites reste ouverte : leur évaluation nécessiterait une validation humaine et constitue une perspective pour de futurs travaux.

5. Conclusion et perspectives

Dans cet article, nous avons proposé deux scores permettant de quantifier les relations de substituabilité et de complémentarité entre *items* à partir de données de sessions d'utilisateurs, avec pour objectif de servir de fondement à la construction de nouvelles métriques d'évaluation hors ligne pour les SRs. Ces métriques ont pour objectif, d'une part, d'évaluer dans quelle mesure les recommandations répondent aux besoins actuels de l'utilisateur et, d'autre part, d'analyser si elles constituent un ensemble cohérent d'*items*, susceptible d'être perçu et utilisé comme un lot. Le recours à des scores de substituabilité et de complémentarité est motivé par l'observation que les interactions observées au sein d'une session traduisent un processus de décision orienté par les objectifs de l'utilisateur, conduisant à l'exploration d'*items* substituables et complémentaires.

La formulation proposée repose uniquement sur les co-occurrences d'*items* et sur la notion de contexte d'un *item* dans une session, ce qui rend le calcul des scores

facilement applicable à des cas d'usage réels à grande échelle, sans nécessiter de modélisation complexe ni d'information externe. Les résultats obtenus sur les jeux de données INCA2 et BundleRec montrent que ces scores permettent d'extraire des relations pertinentes, avec des performances comparables à celles d'approches de la littérature. Nos expérimentations mettent en évidence l'intérêt de la notion de contexte d'un *item* dans une session pour améliorer l'identification des relations de substituabilité, ainsi que l'influence du contexte des sessions sur les relations entre *items*.

Ce travail ouvre plusieurs perspectives de recherche. Une première piste consiste à analyser les relations de substituabilité et de complémentarité qui ne sont pas correctement prédites par les scores proposés afin de mieux comprendre les limites de l'approche. Une seconde piste vise à explorer des stratégies de pondération des contextes d'*items* dans les sessions pour privilégier les contextes les plus pertinents. Plus généralement, cette contribution s'inscrit dans une démarche visant à enrichir l'évaluation des SRs en proposant des métriques orientées vers l'utilité et la cohérence relationnelle, au-delà de la seule précision.

Remerciements

Ce travail a été financé par l'Association Nationale de la Recherche et de la Technologie (ANRT) dans le cadre d'une convention CIFRE (n°2024/1103).

Bibliographie

Adomavicius G., Tuzhilin A., « Context-aware recommender systems », *Recommender systems handbook*, Springer, Boston, MA, p. 217-253, 2010.

AFSSA, Étude Individuelle Nationale des Consommations Alimentaires 2 (INCA 2) (2006–2007), Rapport, Agence française de sécurité sanitaire des aliments, 2009.

Akkoyunlu S., Manfredotti C., Cornuéjols A., Darcel N., Delaere F., « Investigating substitutability of food items in consumption data », *Proceedings of the Second International Workshop on Health Recommender Systems co-located with ACM RecSys*, vol. 5, Como, Italy, August, 2017.

Chen M., Liu P., « Performance Evaluation of Recommender Systems », *International Journal of Performability Engineering*, vol. 13, nº 8, p. 1246, 2017.

Choi J., Park H. Y., « Usage complementarity vs. basket co-occurrence : Discount depth reliance in digitally personalized product recommendations », *Journal of Retailing*, vol. 101, nº 2, p. 177-196, 2025.

Ge M., Delgado-Battenfeld C., Jannach D., « Beyond accuracy : evaluating recommender systems by coverage and serendipity », *Proceedings of the Fourth ACM Conference on Recommender Systems*, Association for Computing Machinery, New York, NY, USA, p. 257–260, 2010. Barcelona, Spain.

He R., McAuley J., « Ups and Downs : Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering », *Proceedings of the 25th International Conference*

on World Wide Web, WWW '16, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, p. 507—517, 2016.

Kaminskas M., Bridge D., « Diversity, Serendipity, Novelty, and Coverage : A Survey and Empirical Analysis of Beyond-Accuracy Objectives in Recommender Systems », *ACM Transactions on Interactive Intelligent Systems*, vol. 7, n° 1, p. 42, 2016.

Kunaver M., Pozrl T., « Diversity in Recommender Systems, A Survey », *Knowledge-Based Systems*, vol. 123, p. 154-162, 02, 2017.

Landsberger H. A., *Hawthorne Revisited : Management and the Worker, Its Critics, and Developments in Human Relations in Industry*, Cornell University Press, Ithaca, N.Y., 1958.

MacKenzie I., Meyer C., Noble S., « How retailers can keep up with consumers », *McKinsey & Company*, vol. 18, n° 1, p. 1-10, 2013.

McNee S. M., Riedl J., Konstan J. A., « Being accurate is not enough : how accuracy metrics have hurt recommender systems », *CHI '06 Extended Abstracts on Human Factors in Computing Systems*, Association for Computing Machinery, New York, NY, USA, p. 1097—1101, 2006. Montréal, Québec, Canada.

Nicholson W., Snyder C., *Demand relationships among goods*, 11th edn, Joe Sabatino, p. 187-205, 2012.

Ravana S. D., Moffat A., « Score Aggregation Techniques in Retrieval Experimentation », *Conferences in Research and Practice in Information Technology Series*, vol. 92, p. 59-67, 01, 2009.

Sun Z., Feng K., Yang J., Fang H., Qu X., Ong Y.-S., Liu W., « Revisiting Bundle Recommendation for Intent-aware Product Bundling », *ACM Transactions on Recommender Systems*, vol. 2, n° 3, p. 34, 2024.

Sun Z., Yang J., Zhang J., Bozzon A., « Exploiting both Vertical and Horizontal Dimensions of Feature Hierarchy for Effective Recommendation », *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017.

Tian Y., Lautz S., Wallis A. O., Lambiotte R., « Extracting complements and substitutes from sales data : a network perspective », *EPJ Data Science*, 2021.

Wang S., Cao L., Wang Y., Sheng Q. Z., Orgun M. A., Lian D., « A Survey on Session-based Recommender Systems », *ACM Computing Surveys*, vol. 54, n° 7, p. 38, 2021.

Wang S., Hu L., Wang Y., Sheng Q. Z., Orgun M., Cao L., « Modeling multi-purpose sessions for next-item recommendations via mixture-channel purpose routing networks », *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, International Joint Conferences on Artificial Intelligence Organization, p. 3771-3777, 7, 2019.

Zangerle E., Bauer C., « Evaluating Recommender Systems : Survey and Framework », *ACM Computing Surveys*, vol. 55, n° 8, p. 38, 2022.

Analyse systématique d'API et scraping adaptatif pour les plateformes de travail numérique

Ayoub FRIHAOUI [1], **Olivier PONS** [1], **Léa LIMA** [2]

Laboratoire Cedric, Conservatoire national des Arts et Métiers
292 Rue Saint-Martin, 75003 Paris, France
firstname.lastname@lecnam.net

RESUME. *Les plateformes de travail numérique structurent l'accès à l'emploi au moyen de systèmes d'information opaques dont les interfaces publiques limitent l'accès aux données nécessaires à l'étude de leurs effets. Cet article traite ce problème sous un angle strictement méthodologique. Les analyses des biais sociaux ou discriminatoires des plateformes sont préliminaires : nous proposons surtout une infrastructure de collecte destinée à rendre de tels audits empiriques possibles. Notre contribution comporte deux volets complémentaires. Nous proposons d'abord un protocole systématique d'analyse des plateformes permettant d'identifier l'architecture d'exposition des données, les filtres imposés, les paramètres cachés et les contraintes de pagination. Nous introduisons ensuite un algorithme de réduction de densité hiérarchique (Hierarchical Density-Reduction Algorithm, HDRA) qui segmente récursivement les requêtes saturées afin d'améliorer l'exhaustivité de la collecte sous contraintes d'API. L'approche est appliquée à quatre plateformes — Sitly, Yoopies, Malt et Care.com — dans trois contextes nationaux. Elle permet de dépasser fortement certaines limites de visibilité, par exemple en récupérant 231999 profils sur Sitly au Brésil malgré un plafond initial de 10000 résultats par requête. Ces résultats montrent l'intérêt d'une démarche reproductible centrée sur la couche de collecte, en amont des futurs audits des systèmes de recommandation et de classement.*

MOTS-CLÉS : *Plateformes numériques de travail , API Web , Collecte de données , Contraintes algorithmiques , Audit algorithmique*

1. Introduction

Les plateformes numériques d'intermédiation du marché du travail jouent aujourd'hui un rôle central dans l'organisation de nombreux secteurs du travail, qu'il s'agisse de garde d'enfants, de services domestiques ou de travail indépendant qualifié (Kalleberg and Vallas, 2017),(Wood *et al.*, 2019),(Howson *et al.*, 2022). Elles ne se contentent pas de faciliter les transactions ; elles structurent l'accès à l'emploi et la visibilité des travailleurs au moyen de systèmes d'information complexes combinant interfaces de recherche, règles de filtrage et algorithmes de classement.

Cependant, le fonctionnement des plateformes reste opaque : si elles constituent des sources riches de micro-données, leurs interfaces publiques sont optimisées pour l'appariement algorithmique et non pour une observation scientifique systématique (Trezza, 2023),(Brenning and Henn, 2023).

Pourtant, comprendre leurs mécanismes constitue un enjeu scientifique majeur pour analyser les formes de segmentation, de hiérarchisation, de biais et de discrimination qu'elles peuvent induire (Vallas and Schor, 2020),(Rosenblat and Stark, 2016).

Cela suppose une analyse statistique à grande échelle des données issues des plateformes. Le premier pas, celui que nous décrivons dans cet article, consiste donc à acquérir ces données de façon automatique dans le respect des règles éthiques et légales.

Cette collecte de données consiste généralement à « scraper ce qui est visible », en formulant des requêtes naïves et en paginant jusqu'à atteindre un plafond. Cette approche néglige souvent des contraintes strictes d'API comme les plafonds fixes de pagination (par exemple 10 000 profils sur la plateforme Sitly au Brésil.) les filtres imposés ou les plages de paramètres cachées. (Lima *et al.*, 2025) montre que ces contraintes ne sont pas neutres : elles produisent des *populations fantômes*, c'est-à-dire des travailleurs présents sur la plateforme mais invisibles aux requêtes standard, générant des biais de sélection silencieux (Brown *et al.*, 2025).

Ce travail s'inscrit dans un projet international de comparaison des plateformes numériques d'intermédiation du marché du travail, des biais et des discriminations qu'elles peuvent induire. Nous nous concentrons toutefois ici sur un enjeu technique central, préalable à toute analyse empirique.

> **Comment peut-on découvrir systématiquement les contraintes d'API propres à une plateforme et concevoir des stratégies de scraping adaptatives permettant de maximiser l'exhaustivité de la collecte, tout en respectant les principes éthiques et les ressources des serveurs ?**

S'appuyant sur des travaux existants en web scraping et en extraction hiérarchique (Chasins *et al.*, 2018) (Latif *et al.*, 2025), notre contribution est avant tout **méthodologique**. L'objectif de cet article n'est pas encore de proposer une analyse substantielle des biais sociaux ou discriminatoires des plateformes, mais de fournir l'infrastructure de collecte nécessaire à de futurs audits empiriques.

Plus précisément, nous proposons une méthodologie en deux volets. Le premier volet consiste en un protocole d'analyse systématique des plateformes, destiné à identifier l'architecture des interfaces de programmation, les filtres imposés, les paramètres cachés et les contraintes de récupération exploitables. Le second volet formalise un algorithme de réduction de densité hiérarchique (*Hierarchical Density-Reduction Algorithm*, HDRA), qui considère le plafond de pagination non comme une simple limite technique, mais comme un indicateur de saturation locale guidant une segmentation récursive de l'espace de recherche.[1]

Nous appliquons cette méthodologie à quatre plateformes analysées en 2025 : Sitly (Brésil), Yoopies (France), Malt (France) et Care.com (États-Unis). Ces cas montrent que les contraintes d'API ne sont pas seulement des détails d'implémentation, mais des objets méthodologiques décisifs pour la qualité, l'exhaustivité et l'interprétabilité des données collectées. En ce sens, notre travail se situe en amont des analyses sociologiques ou algorithmiques proprement dites, en cherchant à réduire le biais de sélection introduit par les stratégies de collecte naïves.

La suite de l'article est organisée comme suit. La section 2 présente l'état de l'art. La section 3 expose les aspects éthiques et légaux. La section 4 décrit le protocole d'analyse des plateformes et formalise HDRA. La section 5 présente les résultats empiriques obtenus sur les différentes plateformes. Enfin, la section 6 revient sur les apports méthodologiques de l'étude et sur ses prolongements possibles pour de futurs audits.

2. Travaux connexes

2.1. *Méthodologies de web scraping*

Le web scraping a évolué du parsing HTML et de l'extraction fondée sur le DOM vers des approches sophistiquées s'appuyant sur des API structurées et des sélecteurs ciblés (Latif *et al.*, 2025), (Brenning and Henn, 2023), (Trezza, 2023). (Latif *et al.*, 2025) montrent que certains attributs récurrents, tels que `data-testid` sur des sites d'e-commerce dynamiques, fournissent des points d'accroche d'extraction robustes ; ils illustrent ainsi les avantages de marqueurs orientés développeurs par rapport à des sélecteurs CSS fragiles. Chasins et al. (Chasins *et al.*, 2018) introduisent *Rousillon*, qui infère des gabarits d'extraction hiérarchiques pour des données web distribuées, en mettant l'accent sur la structure plutôt que sur un parsing ad hoc.

Plus récemment, *ScrapeViz* (Krosnick and Oney, 2024) propose des représentations visuelles pour des macros de scraping multi-étapes. Cependant, ces approches traitent principalement de **la manière d'identifier et d'extraire** du contenu côté client, et non de **la manière de raisonner sur des contraintes d'API côté serveur**—telles que des

1. Par saturation locale, nous entendons une situation dans laquelle une requête atteint la limite maximale de résultats imposée par la plateforme, ce qui signale la nécessité d'un raffinement de l'espace de recherche.

limites de pagination ou des hiérarchies de filtres imposées. (Brenning and Henn, 2023) reconnaissent des contraintes telles que la pagination dans l'acquisition de données géographiques, mais les abordent de façon descriptive plutôt que comme des objets de conception formalisés.

C'est précisément sur ce point que se situe notre contribution. Nous proposons de faire passer les contraintes d'API — plafonds de pagination, hiérarchies de filtres imposés, paramètres cachés, exigences d'état — du statut de détails d'implémentation à celui d'objets méthodologiques de premier ordre. En formalisant l'articulation entre dimensions imposées et dimensions adaptatives, nous reformulons le scraping comme un problème de découverte et d'exploitation de contraintes, cohérent avec une lecture des interfaces comme artefacts à rétroconcevoir (Brown *et al.*, 2025).

2.2. *Travail numérique et données de plateforme*

De nombreux travaux documentent l'essor des plateformes de travail numérique et leur rôle structurant dans le travail à la tâche (Kalleberg and Vallas, 2017),(Vallas and Schor, 2020),(Wood *et al.*, 2019). Les analyses bibliométriques montrent une forte accélération des recherches après 2020, liée à la pandémie et à la généralisation de la coordination algorithmique (Hu and Iahad, 2025). Toutefois, ce champ demeure marqué par une **fragmentation méthodologique**, en particulier concernant l'accès aux données (Silva *et al.*, 2025).

Silva et al. (Silva *et al.*, 2025), à partir d'une revue de 397 études empiriques, montrent que seuls 7% utilisent des données fournies par les plateformes ; la majorité repose sur des déclarations d'utilisateurs ou des données scrapées. Ils soulignent en outre une documentation souvent insuffisante des stratégies d'échantillonnage et des contraintes d'API, limitant l'évaluation de l'exhaustivité et de la reproductibilité. Ce « problème d'accès aux données » est également épistémique, puisqu'il conditionne les questions de recherche formulables (Graham *et al.*, 2017),(Trezza, 2023).

Par ailleurs, les travaux sociologiques mettent en évidence le rôle des plateformes dans la structuration des inégalités d'accès au marché du travail (Lima *et al.*, 2025),(Howson *et al.*, 2022). Tester empiriquement ces mécanismes suppose toutefois l'accès à des populations représentatives, et non à des échantillons tronqués par des plafonds d'API. Notre méthodologie vise précisément à fournir le **substrat de collecte** nécessaire à une analyse robuste.

2.3. *Audit algorithmique et mesure des biais*

L'audit des systèmes de recommandation et de classement connaît un intérêt croissant (Bandy, 2021),(Selbst *et al.*, 2019),(Raji *et al.*, 2020). Des audits valides requièrent des données de référence complètes, condition rarement satisfaite lorsque des contraintes d'API excluent silencieusement certaines sous-populations (Bandy, 2021). L'évaluation hors ligne souffre en outre de biais d'exposition (Carraro

and Bridge, 2022),(Ai *et al.*, 2021), et les méthodes de débiaisage supposent l'accès à l'ensemble des candidats.

Notre objectif est ainsi de produire des **ensembles aussi complets que possible**, afin de distinguer biais algorithmique (classement par la plateforme) et biais d'échantillonnage (visibilité pour le chercheur). Nous positionnons ce travail comme une **couche de données fondationnelle** pour l'étude de la responsabilité et de la transparence des plateformes. Des analyses préliminaires (Lima *et al.*, 2025) montrent que ces jeux de données révèlent des inégalités invisibles dans des échantillons tronqués.

3. Aspect éthique et légaux

Ce travail est mené dans le cadre d'une recherche académique sur des plateformes qui présentent des informations de profil via des interfaces publiques. Conformément aux recommandations antérieures sur le web scraping à des fins de recherche, nous adoptons une approche **d'intérêt public, à faible impact et transparente** (Brown *et al.*, 2025).

Les données collectées comprennent des attributs de profils visibles par les utilisateurs ordinaires (par exemple, nom, descriptions, prix, localisations et avis générés par les utilisateurs *etc.*) et sont anonymisées.

L'objectif principal est la **recherche scientifique** sur la structure des marchés du travail et les mécanismes des plateformes, en mettant l'accent sur la méthodologie de collecte des données et, dans des travaux futurs, sur l'audit des systèmes de recommandation (Brown *et al.*, 2025),(Bandy, 2021). Cette démarche est cohérente avec d'autres recherches utilisant des données web publiquement accessibles pour l'analyse en sciences sociales ou en politiques publiques (Ulbricht, 2020),(Adams, 2022), (Laouenan *et al.*, 2022).

D'un point de vu légal, en France et dans l'Union européenne, le **Règlement Général sur la Protection des Données (RGPD)** encadre strictement la collecte de données personnelles. La recherche scientifique bénéficie d'une base légale spécifique (article 89 du RGPD), autorisant le traitement sous conditions de proportionnalité, de limitation des finalités et de mesures techniques appropriées. Les organismes de recherche peuvent également invoquer l'exception de « fouille de texte et de données » du Code de la propriété intellectuelle (art. L122-5 et L122-5-3), sauf opposition technique explicite (CNIL, 2024). Au Brésil, la **Lei Geral de Proteção de Dados (LGPD)** (D'Oliveira and Cunha, 2024) introduit des protections similaires, autorisant les usages de recherche sous les mêmes principes. Aux États-Unis, le **Computer Fraud and Abuse Act (CFAA)** a longtemps été utilisé pour poursuivre le scraping non autorisé, mais les décisions récentes, notamment *hiQ Labs v. LinkedIn* (2022) suite à l'arrêt *Van Buren v. United States* (2021), ont établi que l'accès à des données publiquement accessibles ne constitue généralement pas un accès « sans autorisation » au sens du CFAA (Sellars, 2018),(Brown *et al.*, 2025).

Cette jurisprudence légitime le scraping de sites publics à des fins de recherche, dès lors que les données sont accessibles et que les pratiques respectent les normes techniques. Notre travail respecte ces principes en limitant la collecte aux attributs de profils publics et en maintenant des cadences de requêtes prudentes.

4. Méthodologie

Pour relever le défi de la récupération de jeux de données complets à partir des « boîtes noires » des plateformes, nous proposons une méthodologie formalisée comprenant deux étapes : (1) un **protocole d'analyse de la plateforme** pour rétro-concevoir les contraintes de l'interface et l'architecture de l'API, et (2) un **algorithme hiérarchique de réduction de densité (HDRA)** pour contourner systématiquement ces contraintes. Nous utilisons *Sitly* (une plateforme de baby-sitting) comme exemple fil conducteur pour illustrer ces concepts.

4.1. *Protocole d'analyse de la plateforme*

Avant toute collecte automatisée, on doit caractériser les barrières techniques de la cible. Ce protocole transforme la plateforme, d'une interface inconnue, en un ensemble de contraintes formelles utilisées pour configurer l'algorithme de scraping.

1. Découverte de l'architecture et de l'interface. La première étape consiste en une exploration manuelle afin de classifier le mode de distribution des données de la plateforme. À l'aide des outils de développement du navigateur (par exemple, l'onglet Network de Chrome), nous distinguons deux architectures principales. Les plateformes **statiques/SSR (Server-Side Rendering)** renvoient du contenu HTML avec une pagination limitée (par exemple, 10 profils par page), nécessitant souvent une analyse du DOM. À l'inverse, les plateformes **dynamiques/SPA (Single Page Applications)**, comme Sitly, interrogent généralement des API internes qui renvoient des données JSON structurées. Ces API prennent souvent en charge des limites de pagination plus élevées (par exemple, 100 profils par page) et contiennent des métadonnées non visibles dans l'interface utilisateur. Au cours de cette phase, nous identifions également les **filtres imposés** (sélection obligatoire d'une ville, d'une catégorie ou d'un code postal) ainsi que les mécanismes de tri. Pour Sitly, nous avons observé que la localisation n'est pas un paramètre de recherche mais un paramètre d'état associé au compte utilisateur (Figure 1).

2. Extraction de modèles et sondage des contraintes. Une fois l'architecture comprise, nous extrayons des modèles canoniques de requêtes (en-têtes, cookies, charges utiles) afin de les rejouer à l'aide d'outils tels que Postman ou Insomnia (Figure 2). Ce sondage manuel nous permet de déterminer la **limite dure de récupération** (τ), définie comme le nombre maximal d'éléments que l'API renvoie pour une requête unique, indépendamment de la profondeur de la pagination. Pour Sitly, le sondage a révélé un plafond global de $\tau = 10\,000$ profils ; pour Care.com, $\tau = 500$. Parallèlement,

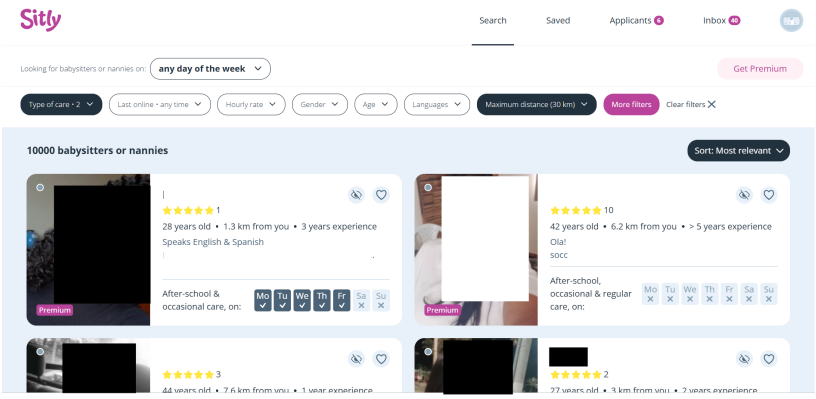


FIGURE 1. *Page de recherche de baby-sitters sur Sitly.com.br illustrant les filtres imposés.*

nous testons l'existence de **plages de paramètres cachées**. Alors que l'interface Sitly limite le filtrage par âge à $[18 - 70+]$, une manipulation directe de l'API via Postman a confirmé que le serveur accepte des âges compris entre 0 et 160 (Figure 3). Cette étape permet de définir les bornes de l'*espace de recherche*.

3. Calibrage des mécanismes de défense. Afin de garantir une collecte éthique et soutenable, nous calibrons empiriquement les limites de taux. Nous utilisons des scripts Python pour rejouer des requêtes à différents intervalles, en surveillant l'apparition d'erreurs HTTP 429, de CAPTCHA ou de blocages temporaires d'adresses IP. Cette phase permet également de vérifier les en-têtes de requête nécessaires (par exemple, User-Agent, jetons d'autorisation). Pour les plateformes étudiées, un délai aléatoire conservateur compris entre 0,6 et 1,0 seconde par requête a permis d'éviter le déclenchement des mécanismes anti-bot.

4.2. *Algorithme hiérarchique de réduction de densité (HDRA)*

Une fois les contraintes cartographiées, nous formalisons le processus d'extraction. L'objectif de HDRA est de récupérer l'ensemble de tous les profils cibles $\mathcal{U}$ malgré la limite dure τ. L'intuition centrale consiste à considérer l'API comme un oracle de requête « boîte noire » et à découper récursivement l'espace de recherche jusqu'à ce que le nombre de résultats passe en dessous de τ, en garantissant qu'aucun profil « fantôme » ne reste caché.

Définition du problème et saturation. Soit S un ensemble de paramètres de recherche (par exemple, $S = \{\text{city : São Paulo, age : 18-24}\}$). Nous interrogeons l'oracle de la plateforme $Q(S)$ afin d'obtenir un ensemble de profils récupérés R_S. La pagi-

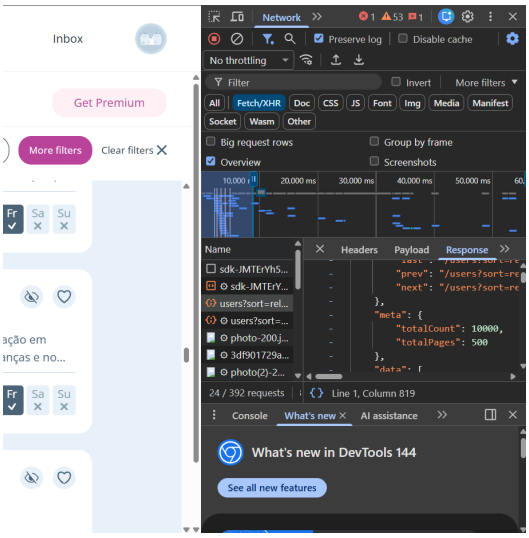


FIGURE 2. *Outils de développement réseau de Chrome inspectant les requêtes API de Sitly.*

nation interne est prise en charge au sein de $Q(S)$ (itération des pages 1 à n) afin de maximiser la récupération, mais la plateforme ne renverra jamais plus de τ profils. Nous définissons un **prédicat de saturation**, $\sigma(S)$, afin de détecter si des données manquent.

– **Type A (compte explicite) :** Si l'API fournit un champ 'totalCount' (par exemple, Sitly), $\sigma(S)$ est vrai si 'totalCount' $\geq \tau$.

– **Type B (censuré) :** Si les comptes sont cachés, $\sigma(S)$ est vrai si $|R_S| = \tau$.

Si $\sigma(S)$ est **vrai**, la requête est « sursaturée », ce qui implique qu'il existe strictement plus de profils que ceux qui ont été renvoyés. Si $\sigma(S)$ est **faux**, nous avons atteint une récupération complète pour ce sous-espace.

Logique de l'algorithme. HDRA opère sur un vecteur ordonné de dimensions de filtrage $\vec{D} = (D_1, D_2, \ldots, D_k)$. L'ordre de ce vecteur n'est pas arbitraire : il est défini à partir du protocole d'analyse présenté en section 4.1. Les dimensions imposées par la plateforme apparaissent en premier, puisqu'elles structurent l'accès même aux résultats ; viennent ensuite les dimensions adaptatives qui présentent à la fois une granularité exploitable et une capacité effective à désaturer les requêtes.[2] Pour Sitly,

2. En pratique, une dimension est retenue lorsqu'elle peut être partitionnée en sous-domaines valides du point de vue de l'API et lorsqu'un raffinement sur cette dimension réduit effectivement la densité locale des résultats.

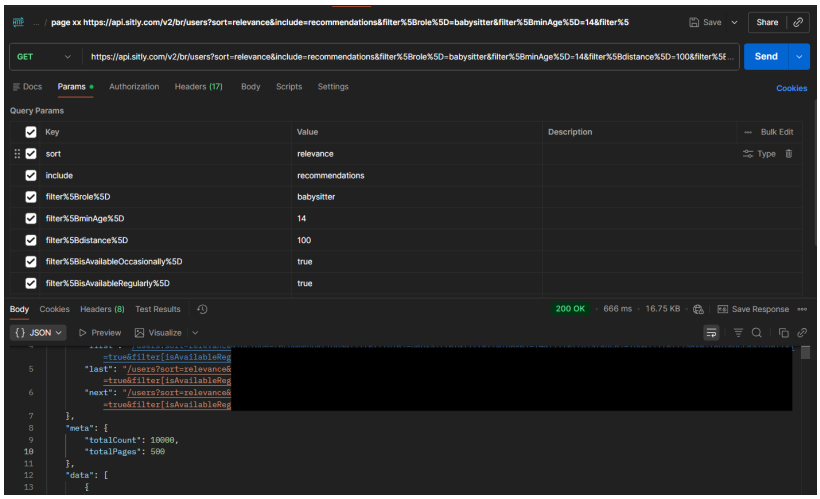


FIGURE 3. *Postman : sondage des points de terminaison de l'API Sitly pour identifier des plages de paramètres cachées.*

par exemple, la localisation imposée par l'état du compte précède la dimension d'âge ; pour Malt et Care.com, la variable prix constitue au contraire une dimension adaptative plus pertinente.

À chaque dimension est associé un opérateur de partition Ψ_i (par exemple, un découpage binaire pour l'âge). L'algorithme procède récursivement : 1. **Interroger** l'API avec les paramètres S. 2. **Vérifier la saturation :** si $\sigma(S)$ est faux, retourner les profils R_S (nœud feuille). 3. **Découper :** si $\sigma(S)$ est vrai, appliquer Ψ_i pour partitionner S en sous-espaces $\{S_1, \ldots, S_m\}$ (par exemple, découper l'âge [18-30] en [18-24] et [25-30]) et réappliquer récursivement l'algorithme. Si la dimension D_i est épuisée (atomique), passer à D_{i+1}.

Dans l'algorithme 1, HARVEST désigne simplement la procédure récursive principale de HDRA : elle prend en entrée un contexte de recherche S, un indice de dimension i, le vecteur $\vec{D}$ et le plafond τ, puis renvoie l'ensemble agrégé des profils récupérés sur le sous-espace considéré.

Complétude et complexité. En supposant que le vecteur initial $\vec{D}$ couvre l'intégralité du domaine de la plateforme (H1), que les filtres opèrent par intersection (H2) et que la population est stable durant la collecte (H3) , l'absence totale de saturation dans les feuilles implique une complétude absolue de la récupération.

Les deux premières hypothèses sous-jacentes à la propriété de complétude ont été vérifiées sur l'ensemble des plateformes étudiées : H1 par sondage exhaustif des

Algorithme 1: Hierarchical Density-Reduction Algorithm (HDRA)

Input : Search Context S, Dim Index i, Vector $\vec{D}$, Cap τ
Output : Set of unique profiles R_{total}

$(R_S, \kappa) \leftarrow$ `QueryAPI`(S);
`// Fetch profiles & metadata`
if $\neg$`IsSaturated`$(\kappa, |R_S|, \tau)$ **then**
  return R_S ; `// Base case: complete`
if $i >$ *length*$(\vec{D})$ **then**
  return R_S ; `// Max depth reached`
$Subspaces \leftarrow \Psi_i$.`Partition`$(S)$; `// e.g., Split Age range`
if $Subspaces \neq \emptyset$ **then**
  $R_{total} \leftarrow \emptyset$;
  foreach $S_j \in Subspaces$ **do**
    $R_{total} \leftarrow R_{total} \cup$ `Harvest`$(S_j, i, \vec{D}, \tau)$;
  return R_{total};
else
  `// Dim atomic, move to next filter`
  return `Harvest` (S, i+1, $\vec{D}, \tau$)

bornes de paramètres, H2 par test empirique de monotonie lors de la phase d'analyse. H3 correspond à une hypothèse de type snapshot, raisonnable au regard de la durée finie des campagnes de collecte, supposée courte relativement au rythme de création, suppression ou modification des profils.

La complexité en nombre de requêtes dépend de la densité, et est gouvernée par :

$$C_{HDRA} \approx O\left(\frac{N}{\tau} \cdot \beta \cdot \gamma\right) \quad [1]$$

où :

– N est la taille totale de la population et τ est le plafond imposé par la plateforme.

– $\beta \geq 1$ est le **coefficient de redondance**, représentant le recouvrement dû à des requêtes non disjointes (par exemple, des rayons géographiques qui se recouvrent).

– γ représente le **déséquilibre de distribution**. Les distributions uniformes minimisent les découpages ($\gamma \approx 1$), tandis que des distributions fortement déséquilibrées (par exemple, 90% des utilisateurs dans une seule ville) imposent une récursion profonde ($\gamma > 1$).

Pour Sitly, $\beta \approx 1.2$ en raison du recouvrement radial, tandis que γ était significatif en raison de la densité à São Paulo.

4.3. *Implémentation du pipeline et limites*

L'implémentation comprend un **gestionnaire de session** pour gérer les exigences avec état (comme le paramétrage de la localisation dans Sitly) ainsi qu'un **ordonnanceur** pour faire respecter les limites de taux. De manière cruciale, le pipeline journalise de nombreuses **métadonnées** (horodatages, comptes totaux, tailles de réponse) en parallèle des données de profils afin de vérifier l'intégrité des données. Il est à noter que le post-traitement (déduplication et nettoyage) est effectué strictement après la collecte et se situe en dehors du périmètre de cet article. Dans cette architecture, le noyau de HDRA (détection de saturation, logique de partition, orchestration récursive, journalisation et déduplication) est réutilisable d'une plateforme à l'autre, tandis que les connecteurs spécifiques - authentification, format des requêtes, paramètres imposés, sémantique des métadonnées comme `totalCount`, et choix initial des dimensions $\vec{D}$ — doivent être ajustés au cas par cas lors du protocole d'analyse.

Limites. Bien qu'efficace, cette méthodologie présente des modes d'échec spécifiques :

– **Limitation côté serveur :** si les limites sont strictes (par exemple, < 100 requêtes/heure), une récursion profonde devient irréalisable sans rotation d'adresses IP distribuées.

– **Filtrage non monotone :** HDRA suppose que l'ajout d'un filtre réduit le nombre de résultats (intersection). Si une plateforme utilise une logique d'union (OR), les découpages ne réduiront pas la densité.

– **Dimensions non orthogonales :** si le filtrage selon D_1 modifie les options valides pour D_2 (par exemple, des catégories qui modifient les plages de prix disponibles), l'opérateur de partition Ψ doit être conscient de l'état.

– **Churn temporel :** les données sont collectées sur une fenêtre (par exemple, 42 heures), ce qui crée une sémantique de « snapshot » plutôt qu'une vue instantanée.

5. Résultats empiriques

Nous présentons maintenant des résultats empiriques issus des quatre plateformes, en nous concentrant sur la complétude de la récupération, la gestion des contraintes et l'efficacité à haut niveau. Nous détaillons en particulier le comportement de l'API concernant la pagination et la manière dont les dimensions imposées ont été gérées.

5.1. *Vue d'ensemble des plateformes*

Le tableau 1 résume les principales caractéristiques techniques pertinentes pour la conception de HDRA. Le scraping a été réalisé en 2025. Il est à noter que les limites de paramètres (par exemple, l'intervalle d'âge [0–160] pour Sitly) ont été découvertes avant le scraping à l'aide des scripts de sondage automatisés décrits dans

la Méthodologie, qui ont testé de manière incrémentale les valeurs limites jusqu'à ce que l'API renvoie des erreurs ou des ensembles vides.

Comme point de comparaison minimal, nous considérons ici une stratégie naïve consistant à paginer exhaustivement la requête la plus large autorisée, sans re-segmentation de l'espace de recherche : sur les plateformes dotées d'un plafond dur (Sitly, Malt, Care.com), cette stratégie reste mécaniquement bornée par τ, tandis que le gain rapporté dans le tableau 1 provient précisément de la subdivision adaptative mise en œuvre par HDRA ; à l'inverse, sur Yoopies, où aucun plafond global dur n'a été observé, l'itération paginée suffit.

TABLEAU 1. *Caractéristiques des plateformes et profils de contraintes*

Plateforme	Type d'API	Dimension(s) imposée(s)	Dimension(s) adaptative(s)	Plafond τ	Profils récupérés	Gain
Sitly Brazil	SPA (JSON)	Ville (localisation du compte)	Âge	10,000	231,999 uniques	x23.2
Yoopies France	SPA (JSON)	Catégorie (garde d'enfants, …)	Rayon, ville	Aucun	49,205	-
Malt France	Hybride	Mot-clé de recherche, code postal	Prix	300	>66,000	x220
Care.com USA	SPA (JSON)	Catégorie, code postal	Prix	500	8,417 (NYC)	x16.8

5.2. *Sitly Brazil : récupération tenant compte des contraintes*

5.2.1. *Stratégie de sélection des villes*

Les profils Sitly sont concentrés dans les zones urbaines, mais la plateforme impose la localisation via les paramètres du compte plutôt que par un filtre de recherche. Afin de cartographier le pays, nous avons employé une approche fondée sur des centroïdes. Nous définissons un **centroïde** comme une paire de coordonnées latitude/longitude spécifique servant de point central pour une recherche géographique.

1) Nous avons compilé une liste de 382 villes brésiliennes à l'aide de données géographiques externes.

2) Nous les avons triées par densité de population.

3) Nous avons sélectionné les **161 premières villes** pour servir de centroïdes.

Pour chaque centroïde, nous avons configuré le **rayon** de recherche, c'est-à-dire, la distance maximale à partir du point central à partir de laquelle les profils sont inclus. Nous avons utilisé un rayon de 127 km, une limite identifiée par sondage de l'API comme étant la portée maximale effective avant que le serveur n'ignore le paramètre.

5.2.2. *Exécution et pagination*

Pour chaque centroïde, le protocole s'est déroulé comme suit :

– **Mise à jour de l'état :** nous nous sommes authentifiés et avons invoqué le point de terminaison de l'API « set location » afin de mettre à jour les coordonnées du compte du scraper vers le centroïde courant (par exemple : passage de São Paulo à Rio de Janeiro).

– **Sondage :** nous avons envoyé une requête de recherche afin de lire la métadonnée `totalCount`.

– **Pagination (sous le plafond) :** si `totalCount` $< 10{,}000$, nous avons récupéré l'ensemble des profils. Comme l'API Sitly limite les réponses à **100 profils par page**, cela a nécessité d'itérer de `page=1` jusqu'à `page=100` (ou $\lceil \texttt{totalCount}/100 \rceil$).

– **Saturation (au-dessus du plafond) :** si `totalCount` atteignait la limite dure de 10,000, nous avons marqué la requête comme saturée. Au lieu de paginer immédiatement (ce qui ferait perdre les profils au-delà du 10 000e), nous avons invoqué HDRA pour découper la requête à l'aide de la dimension Âge.

Cette stratégie réduit la **redondance** (moins de rayons qui se recouvrent que dans des grilles arbitrairement denses) et le **gaspillage** (en évitant des requêtes sur des zones rurales presque vides).

5.2.3. *HDRA basé sur l'âge à São Paulo*

São Paulo était la région la plus saturée. Une seule requête pour « tous les âges » au centroïde de São Paulo a renvoyé `totalCount` $= 10{,}000$, confirmant que le plafond de la plateforme était atteint. HDRA a découpé récursivement les intervalles d'âge en utilisant les limites de l'API découvertes [0–160] :

1) Requête Âge [0–160] $\rightarrow$ saturée. Découpage.

2) Requête Âge [0–80] et [81–160].

3) La branche [0–80] est restée saturée, déclenchant de nouveaux découpages en [0–40], [41–80], puis jusqu'à des bandes étroites comme [24–25].

Résultats :

– L'agrégation a posteriori a révélé **61,578** apparitions brutes de profils pour São Paulo.

– Après suppression des doublons (dus au recouvrement des rayons de villes ou à des cas limites sur les bornes d'âge), nous avons obtenu **51,371** profils uniques.

– Le coefficient de redondance était $\beta \approx 1.20$, ce qui signifie que seulement 20% des profils récupérés étaient des doublons, indiquant une efficacité élevée.

5.2.4. *Couverture nationale*

L'extension de cette approche aux 161 centroïdes a permis d'obtenir **231,999 profils uniques** à l'échelle du Brésil. Pour la majorité des villes de plus petite taille, le sondage initial a renvoyé un compte inférieur à 10,000, permettant une récupération complète par simple pagination, sans segmentation supplémentaire.

5.3. *Yoopies : recherche imposée par catégorie*

Yoopies.fr impose des catégories de premier niveau via la structure des URL (par exemple, `/childcare-sitters`, `/housekeeping`), faisant de la catégorie la

première dimension imposée D_1. Lors de notre collecte d'avril 2025 :

– Nous avons itéré sur les catégories imposées.

– Nous avons appliqué un paramétrage de **rayon** maximal (effectivement national) afin de capturer l'ensemble des profils en France. Il faut noter que la limite de 30 km imposée par le curseur de l'interface graphique n'est qu'une restriction de la couche de présentation : l'analyse a révélé que l'API ne bride aucunement la distance de recherche.

– L'API renvoyait du JSON paginé standard (100 profils par page). Contrairement à Sitly, nous n'avons rencontré aucun plafond global dur (par exemple, 10,000) bloquant la pagination ; l'itération sur les pages a suffi pour récupérer l'ensemble des **49,205** profils.

5.4. *Malt et Care.com : segmentation basée sur le prix*

Malt.fr impose une hiérarchie stricte : il faut d'abord sélectionner un **mot-clé de recherche** (par exemple, « Developer ») ainsi qu'une **localisation** (au niveau de la ville ou du pays). L'API limite les résultats à $\tau = 300$. Afin de maximiser la récupération :

– Nous avons défini $D_{search\_keyword}$ et $D_{location}$ comme contexte de base.

– Nous avons utilisé le **prix** (tarif horaire) comme dimension adaptative D_{price}.

– Lorsqu'une requête renvoyait 300 profils, nous découpions l'intervalle de prix (par exemple, 0–50 $\rightarrow$ 0–25, 26–50) jusqu'à ce que les comptes passent en dessous de 300. (l'heuristique fondée sur le prix était possible parce que la plateforme affiche dans l'interface utilisateur la distribution du nombre de freelances par histogramme de prix).

Cette méthode a permis d'obtenir plus de **66,000** profils de freelances. Comme les tranches de prix sont disjointes (un utilisateur est soit à 20 $, soit à 21 $, jamais les deux), la redondance est proche de la perfection ($\beta \approx 1$).

Care.com (USA) impose de manière similaire une **catégorie** (par exemple, « Babysitter ») et un ancrage par **code postal**, et limite les résultats à $\tau = 500$.

– Nous avons ciblé la catégorie souhaitée ainsi qu'un code postal à forte densité à New York.

– Nous avons utilisé le prix/heure comme dimension de découpage.

– À partir d'une requête de base saturée unique (500 résultats), la segmentation par le prix a permis d'étendre la récupération à **8,417** profils distincts pour ce seul emplacement.

6. Conclusion et perspectives

Le travail présenté dans cet article se concentre sur la **couche de collecte** des données. Les développements complémentaires relatifs au traitement des profils extraits

n'entrent donc pas dans le périmètre de cette contribution. Notre apport principal est méthodologique : nous proposons une démarche articulant un protocole d'analyse de plateforme et l'algorithme *Hierarchical Density-Reduction Algorithm* (HDRA), afin d'identifier et d'exploiter systématiquement les contraintes d'accès aux données.

Cette approche permet de traiter comme objets méthodologiques de premier ordre des éléments souvent laissés à l'arrière-plan, tels que les plafonds de pagination, les filtres imposés, les paramètres cachés ou encore les mécanismes avec état. En ce sens, l'article ne prétend pas encore produire une analyse substantielle des biais sociaux ou discriminatoires des plateformes ; il fournit plutôt l'infrastructure de collecte nécessaire à de futurs audits empiriques. Les résultats obtenus sur plusieurs plateformes et dans plusieurs contextes nationaux montrent qu'un tel cadrage permet d'améliorer fortement l'exhaustivité de la collecte, tout en maintenant des conditions de requêtage prudentes.

À ce stade, le vecteur de dimensions $(\vec{\mathcal{D}})$ et les opérateurs de partition (Ψ_i) sont définis manuellement. Des travaux futurs pourraient exploiter les sondages initiaux pour automatiser l'apprentissage des ordonnancements de dimensions et optimiser les stratégies de découpage, l'objectif étant de minimiser le nombre total de requêtes au regard des schémas de saturation observés.

Le traitement des données extraites, exploré dans des travaux complémentaires en sociologie, a révélé des résultats parfois contre-intuitifs. Il apparaît notamment que, si l'interface semble favoriser une sélection par les prix, un tarif élevé couplé à la maîtrise du français comme langue seconde (indicateur indirect d'origine migratoire) corrèle positivement avec l'accès à l'emploi. Ces travaux ont également fait émerger de nouveaux besoins, en particulier la nécessité d'une **analyse longitudinale** par une surveillance et un audit réguliers.

Pour y répondre, nous prévoyons de mobiliser nos stratégies de collecte pour organiser des campagnes périodiques sur des contextes fixes (par exemple, « babysitter à São Paulo »). L'objectif est de constituer des séries temporelles permettant de mesurer la stabilité et le renouvellement des classements, tout en appliquant des métriques d'équité et des cadres d'audit reconnus.

Bibliographie

Adams N. N., « Scraping'Reddit posts for academic research ? Addressing some blurred lines of consent in growing internet-based research trend during the time of Covid-19 », *International journal of social research methodology*, vol. 27, n° 1, p. 47-62, 2022.

Ai Q., Yang T., Wang H., Mao J., « Unbiased learning to rank : Online or offline ? », *ACM Transactions on Information Systems (TOIS)*, vol. 39, n° 2, p. 1-29, 2021.

Bandy J., « Problematic Machine Behavior : A Systematic Literature Review of Algorithm Audits », , vol. 5, 1, p. 1-34, 2021.

Brenning A., Henn S., « Web scraping : a promising tool for geographic data acquisition », *arXiv preprint arXiv :2305.19893*, 2023.

Brown M. A., Gruen A., Maldoff G., Messing S., Sanderson Z., Zimmer M., « Web scraping for research : Legal, ethical, institutional, and scientific considerations », *Big Data & Society*, vol. 12, n° 4, p. 20539517251381686, 2025.

Carraro D., Bridge D., « A sampling approach to Debiasing the offline evaluation of recommender systems », *Journal of Intelligent Information Systems*, vol. 58, n° 2, p. 311-336, 2022.

Chasins S. E., Mueller M., Bodik R., « Rousillon : Scraping distributed hierarchical web data », *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, p. 963-975, 2018.

CNIL, « Recommandations sur l'intérêt légitime et la collecte de données par moissonnage (web scraping) », , Commission Nationale de l'Informatique et des Libertés, 2024. Consulté le 16 février 2026.

D'Oliveira N. P. C., Cunha F. J. A. P., « Brazilian General Data Protection Law (LGPD) : the relationship between information policy and information regime », vol. 22, SciELO Brasil, p. e024015, 2024.

Graham M., Lehdonvirta V., Wood A., Barnard H., Hjorth I., Peter Simon D., « The risks and rewards of online gig work at the global margins », *Oxford Internet Institute*, 2017.

Howson K., Johnston H., Cole M., Ferrari F., Ustek-Spilda F., Graham M., « Unpaid labour and territorial extraction in digital value networks », 10, 2022.

Hu L., Iahad N. A., « A Decade of Gig Economy Research (2014–2025) : A Bibliometric and Scientific Mapping of Digital Labor Scholarship », *Journal of Information Systems and Technology Management*, 2025.

Kalleberg A. L., Vallas S. P., « Probing Precarious Work : Theory, Research, and Politics », *Precarious Work*, Emerald Publishing Limited, 12, 2017.

Krosnick R., Oney S., « Scrapeviz : Hierarchical representations for web scraping macros », *2024 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, IEEE, p. 300-305, 2024.

Laouenan M., Chapelle G., Deschamps P., Glover D., Lambin X., Seshie M., Grisolia P., Alaye S., Henry C., Les discriminations en raison du genre et de l'origine supposée sur deux plateformes collaboratives, PhD thesis, LIEPP ; Défenseur des droits, 2022.

Latif A., Wildah S. K., Agustiani S., Juningsih E. H., « Implementation of a Data-Testid Attribute-Based Web Scraping Method for Accommodation Data Extraction from a Dynamic E-Commerce Website (Case Study : Traveloka) », *Journal of Artificial Intelligence and Engineering Applications (JAIEA)*, 2025.

Lima L., Guimarães N. A., Pons O., « HOW PLATFORMS STRUCTURE UNEQUAL ACCESS TO EMPLOYMENT OPPORTUNITIES IN PAID DOMESTIC WORK. A COMPARISON OF TWO ONLINE MARKETPLACES IN FRANCE AND BRAZIL 1 », 2025.

Raji I. D., Gebru T., Mitchell M., Buolamwini J., Lee J., Denton R., « Saving face : Investigating the ethical concerns of facial recognition auditing », *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, p. 145-151, 2020.

Rosenblat A., Stark L., « Algorithmic labor and information asymmetries : A case study of Uber's drivers », *International journal of communication*, vol. 10, p. 27-27, 2016.

Selbst A. D., Boyd D., Friedler S. A., Venkatasubramanian S., Vertesi J., « Fairness and Abstraction in Sociotechnical Systems », *Proceedings of the conference on fairness, accountability, and transparency*, p. 59-68, 2019.

Sellars A., « Twenty years of web scraping and the computer fraud and abuse act », *BUJ Sci. & Tech. L.*, vol. 24, p. 372, 2018.

Silva V., Susha I., Hoekman J., Frenken K., « How Data Access Shapes Research into Labour Platforms », *SSRN Working Paper*, 2025. Available at SSRN : https://ssrn.com/abstract=5213743.

Trezza D., « To scrape or not to scrape, this is dilemma. The post-API scenario and implications on digital research », *Frontiers in sociology*, vol. 8, p. 1145038, 2023.

Ulbricht L., « Scraping the demos. Digitalization, web scraping and the democratic project », *Democratization*, vol. 27, n° 3, p. 426-442, 2020.

Vallas S., Schor J. B., « What do platforms do ? Understanding the gig economy », *Annual review of sociology*, vol. 46, n° 1, p. 273-294, 2020.

Wood A. J., Graham M., Lehdonvirta V., Hjorth I., « Good Gig, Bad Gig : Autonomy and Algorithmic Control in the Global Gig Economy », *Work, Employment and Society*, vol. 33, n° 1, p. 56-75, 2019. PMID : 30886460.

Modélisation et visualisation de trajectoires territoriales

Yunji ZHANG[1,3], Sebastien LABORIE[1], Philippe ROOSE[1], Franck RAVAT[2]

1. Universite de Pau et des Pays de l'Adour, E2S UPPA, LIUPPA, Anglet, France
{yunji.zhang, sebastien.laborie, philippe.roose}@univ-pau.fr

2. Université Toulouse 1 Capitole, IRIT, Toulouse, France
franck.ravat@irit.fr
3. Domolandes Digital Lab, Technopole Domolandes, Saint-Geours-de-Maremne, France

RESUME. *Cet article est une synthèse de l'article original : Zhang, Y., Laborie, S., & Roose, P. (2026). Modeling and Visualizing Territorial Trajectories. In Proceedings of the 20th International Conference on Research Challenges in Information Science (RCIS 2026).*

MOTS-CLÉS : *trajectoire territoriale, analyse spatio-temporelle, visualisation analytique, développement territorial, bien-vivre et bien-vieillir*

1. Introduction et problématique

Le développement territorial résulte d'interactions complexes (Torre, 2025). Dans la planification du bien-vivre et du bien-vieillir, les décideurs ont besoin d'outils pour interpréter des données spatio-temporelles multi-thématiques hétérogènes. Les évaluations existantes, souvent mono-thématiques (Stern and Kissinger, 2025), offrent peu de supports visuels pour des comparaisons longitudinales entre granularités différentes. Le concept de trajectoire offre un cadre pertinent pour analyser l'évolution d'une entité à travers le temps, l'espace et différentes dimensions thématiques. Cependant, les travaux existants sur les modèles de trajectoires (Parent, 2013) se concentrent principalement sur les objets mobiles ou les parcours de vie individuels (Gensel, 2020), tandis que l'évolution des territoires en tant qu'entités spatio-temporelles reste encore peu modélisée. Cet article répond à la question suivante : *Comment modéliser des trajectoires*

territoriales à des fins d'analyse spatio-temporelle multi-thématiques sur des jeux de données hétérogènes ?

2. Modèle de trajectoire territoriale

Nous introduisons le concept de **trajectoire territoriale** comme une séquence ordonnée d'états territoriaux. Chaque état décrit une zone spatiale à une période donnée selon une ou plusieurs thématiques. L'originalité de ce modèle réside dans l'intégration, au sein d'une même structure, des dimensions spatiale, temporelle et thématique, ainsi que des attributs et indicateurs associés. Il fournit ainsi un cadre commun pour aligner des sources hétérogènes, comparer des observations issues de granularités variées et soutenir des visualisations adaptées. Nous distinguons deux types de trajectoires : les trajectoires *naturelles*, fondées sur l'ordre chronologique, et les trajectoires *personnalisées*, définies par l'analyste.

3. Implémentation et validation

Les concepts ont été implantés dans un prototype de visualisation dynamique et interactive de trajectoires territoriales (Power BI), mettant en œuvre l'instanciation d'états pilotée par les métadonnées. La validation a été réalisée sur 48 sources hétérogènes (INSEE, data.gouv.fr), démontrant la capacité du modèle à intégrer des indicateurs multi-thématiques (éducation, emploi, revenus) avec des granularités variées.

4. Conclusion

Cet article propose un cadre conceptuel original, les **trajectoires territoriales**, pour modéliser l'évolution des territoires comme une séquence d'états à périmètres explicites. Ce cadre permet d'intégrer et comparer des indicateurs multi-thématiques issus de données hétérogènes. Le prototype montre le potentiel de l'approche pour l'aide à la décision territoriale. Les travaux futurs viseront à enrichir le prototype par des fonctionnalités de visualisation et d'interaction plus avancées.

Bibliographie

Gensel J. e. a., « Un modèle multi points de vue pour représenter les trajectoires de vie », *CIST2020*, p. 173-177, Nov, 2020.

Parent C. e. a., « Semantic trajectories modeling and analysis », *ACM Comput. Surv.*, vol. 45, n° 4, p. 1-32, 2013.

Stern A., Kissinger M., « A multi-perspective framework for assessing urban wellbeing, development, and sustainability », *Habitat International*, vol. 156, p. 103269, 2025.

Torre A., « Territorial development : towards a dynamic and innovative understanding », *Regional Studies*, vol. 59, p. 2465657, 2025.

Méthodes de réduction de données pour la régression : une approche multi-objectifs

Vlada Stegarescu[1,2], Franck Ravat[1], Jiefu Song[1], Leonidas Papastamatis[2], Benoit Baurens[2]

1. Université Toulouse Capitole, IRIT, CNRS (UMR 5505)
2, rue du Doyen-Gabriel-Marty, 31042 Toulouse cedex 9
Vlada.Stegarescu@irit.fr, Franck.Ravat@irit.fr, Jiefu.Song@irit.fr

2. Akkodis Research, 7 Bd Henri Ziegler, 31700 Blagnac
Vlada.Stegarescu@akkodis.com, Benoit.Baurens@akkodis.com

RESUME. *Ce texte est le résumé de l'article : Stegarescu, V., Song, J., Papastamatis, L., Baurens, B. Data Reduction Methods for Regression: A multi-Objective Perspective In proceedings of the 20th International Conference on Research Challenges in Information Science (RCIS 2026)*

MOTS-CLÉS : *Réduction de données, Régression, Sélection Multi-objectifs, Sélection contextuelle*

La régression est une tâche centrale de l'analyse prédictive, visant à modéliser la relation entre des variables explicatives et une variable cible continue. Dans ce cadre, la réduction de données est couramment utilisée pour limiter le surapprentissage, améliorer la stabilité des estimations et réduire le coût de calcul (Kao *et al.*, 2016). Cependant, le choix d'une méthode de réduction constitue un problème complexe, car son efficacité dépend fortement des **caractéristiques du jeu de données** — telles que la linéarité des relations, la distribution des variables, leur type, la dimensionnalité ou encore la taille de l'échantillon — ainsi que de plusieurs **objectifs** souvent conflictuels, notamment la performance prédictive, l'interprétabilité, le taux de réduction et le coût computationnel (Stegarescu *et al.*, 2026). L'importance relative de ces objectifs peut varier selon l'utilisateur et la finalité de l'analyse. Dans cette perspective, la problématique consiste à déterminer, pour un jeu de données donné, quelle méthode de réduction privilégier afin de soutenir efficacement une tâche de régression, en tenant compte à la fois des caractéristiques des données et des préférences de l'utilisateur quant à l'importance relative des différents objectifs. En l'absence d'une approche structurée permettant d'intégrer explicitement ces éléments, et compte tenu de la diversité des

métriques d'évaluation en régression (Chatterjee and Hadi, 2015), la comparaison et la sélection des méthodes de réduction restent délicates.

Afin de répondre à cette problématique, nous formulons la sélection de méthodes de réduction comme un problème de décision multi-objectifs contraint par les caractéristiques des jeux de données qui déterminent le contexte. Pour un contexte donné, chaque méthode est appliquée et décrite par un vecteur de résultats selon les objectifs considérés et ces vecteurs sont ordonnées en fonction d'un ordre de priorité défini par l'utilisateur. Sur cette base, nous introduisons un système d'aide à la décision structuré en deux phases complémentaires. Dans une phase hors ligne, des jeux de données synthétiques sont générés de manière contrôlée afin de couvrir différents contextes. Pour chaque contexte, les méthodes de réduction compatibles sont évaluées à l'aide de plusieurs modèles de régression, et leurs résultats sont représentés sous forme de vecteurs multi-objectifs. L'analyse de Pareto permet alors d'identifier, pour chaque contexte, les compromis pertinents entre les objectifs. Dans une phase en ligne, cette base de connaissances est exploitée pour recommander des méthodes adaptées à un nouveau jeu de données. À partir des caractéristiques des données et de l'ordre de priorité spécifié par l'utilisateur, les méthodes issues du front de Pareto sont classées de manière lexicographique, ce qui permet de fournir une recommandation cohérente, explicite et adaptée aux préférences.

Pour l'évaluation expérimentale, la validation du système est réalisée sur des jeux de données réels issus de notebooks Kaggle. Deux scénarios sont considérés. Dans un premier scénario, où les priorités sont alignées avec les pratiques usuelles (performance en tête), on observe que la première recommandation du système correspond aux choix de réduction du notebook, ou constitue, une alternative empiriquement plus performante. Dans un second scénario, les priorités sont modifiées afin d'évaluer la sensibilité du système : les recommandations évoluent de manière cohérente, en privilégiant par exemple des méthodes plus simples lorsque le coût computationnel devient prioritaire.

En conclusion, ce travail propose un cadre transparent et reproductible pour la sélection de méthodes de réduction en régression, fondé sur une caractérisation empirique et une prise en compte explicite des priorités utilisateur. Il ouvre des perspectives vers un enrichissement de la description des contextes et une extension à d'autres tâches d'apprentissage, ainsi que vers des mécanismes adaptatifs de recommandation.

Bibliographie

Chatterjee S., Hadi A., *Regression Analysis by Example*, 5th edn, Wiley, Hoboken, 2015.

Kao Y.-C., Hsu Y.-S., Lo Y.-C., Lee Y.-H., « A Survey About Prediction-Based Data Reduction in Wireless Sensor Networks », *2016 IEEE International Conference on Communications (ICC)*, IEEE, Piscataway, p. 1-6, 2016.

Stegarescu V., Ravat F., Song J., Papastamatis L., Baurens B., « Enabling Context-Aware Data Reductions », *Advances in Intelligent Data Analysis XXIV (IDA 2026)*, vol. 16513 of *Lecture Notes in Computer Science*, Springer, Cham, 2026.

Vers l'automatisation de la gestion du cycle de vie des jumeaux numériques

Gwendal Beaumont[1,2], Antoine Beugnard[1], Salvador Martínez[1], Christelle Urtado[2], Sylvain Vauttier[2]

1. IMT Atlantique, Lab-STICC (UMR 6285)
Technopôle Brest-Iroise CS 83818
29238 Brest Cedex 3, France
{gwendal.beaumont,antoine.beugnard,salvador.martinez}@imt-atlantique.fr

2. IMT Mines Alès, SyCoIA
6 Av. de Clavières
30100 Alès, France
{christelle.urtado,sylvain.vauttier}@mines-ales.fr

RESUME. *Cet article est une synthèse de l'article : Beaumont, G., Beugnard, A., Martínez, S., Urtado, C., Vauttier, S. Towards Automating the Life Cycle Management of Digital Twins. In: Bork, D., Lukyanenko, R., Sadiq, S., Bellatreche, L., Pastor, O. (eds) Conceptual Modeling. ER 2025. Lecture Notes in Computer Science, vol 16189. Springer, Cham. https://doi.org/10.1007/978-3-032-08623-5_22*

MOTS-CLÉS : *Jumeaux numériques, Ingénierie dirigée par les modèles, DevOps, GitOps*

Le terme *jumeau numérique* (JN) fait référence à une entité virtuelle qui représente, copie et peut contrôler un système, un processus ou un objet donné, appelé système de référence (RS). La nature du RS, ainsi que les objectifs du JN peuvent êtres multiples. Les JN ont pour avantage de permettre, entre autres, la visualisation, la simulation ou la prédiction de l'état du RS. L'intérêt pour cette technologie ne cesse de croître. Des propositions récentes visent à standardiser le développement des JN (Aissat *et al.*, 2025), pourtant notre expérience montre qu'ils sont généralement créés de manière *ad hoc* et sans réelle prise en compte de leur cycle de vie (Beaumont *et al.*, 2024). Ces lacunes représentent des freins à la modularité, la réutilisation et la gestion de ces JN. Cet article propose l'établissement d'un cycle de vie opérationnel du JN ainsi qu'un processus de génération automatique d'artefacts GitOps permettant l'automatisation dudit cycle.

Dans un premier temps, à partir de notre expérience et d'un état de l'art, nous formalisons un ensemble d'opérations élémentaires caractérisant la phase opérationnelle du cycle de vie des JN. Cette base, conçue pour être extensible selon les besoins spécifiques, se compose des opérations suivantes : *start*, *configure*, *connect*, *disconnect*, *synchronize*, *clone*, *save*, *stop*, *archive*, *restore* et *delete*. Ces opérations intègrent des arguments permettant de configurer leurs points de variabilité (voir Figure 1). Par ailleurs, nous observons que l'exécution de ces opérations révèle une interdépendance temporelle. Par exemple, il paraît inconcevable d'arrêter l'instance d'un JN sans l'avoir démarré au préalable. Ainsi, la Figure 1 présente une mise en œuvre concrète d'enchaînement de ces opérations. Enfin, un exemple est proposé au travers du cas d'utilisation sur une usine d'apprentissage *Fischertechnik*.

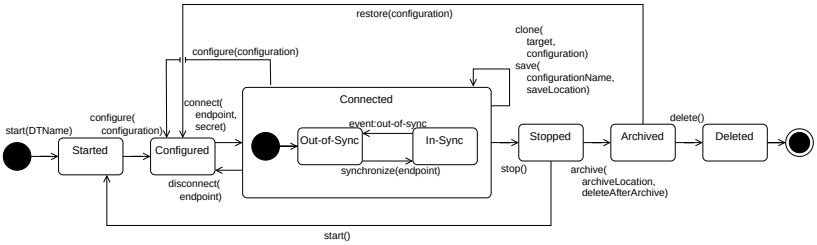


FIGURE 1. *Cycle de vie opérationnel de JN sous la forme d'une machine à états*

Dans un second temps, nous proposons l'automatisation de la gestion de ces opérations grâce à un processus intégrant des techniques d'ingénierie dirigée par les modèles appliquées à une approche *GitOps*. GitOps est l'application des principes *DevOps* à des outils basés sur le logiciel de gestion de version *Git*. Concrètement, après avoir défini un modèle conceptuel de l'approche GitOps, nous proposons un processus en quatre étapes permettant la génération d'artefacts pouvant être directement utilisés par les outils GitOps.

Étape 1 : Un modeleur de JN construit et produit des modèles du JN. Cette modélisation peut se réaliser via divers langages comme *Digital Twin Definition Language*[1] (DTDL) ou *Eclipse Vorto*[2].

Étape 2 : En parallèle, un ingénieur DevOps / GitOps définit l'infrastructure de la plateforme sur laquelle sera installé le futur JN. Cela inclut le choix de plusieurs outils GitOps, dont un outil d'orchestration et un outil de conteneurisation. Ces derniers représentent les éléments cibles d'une plateforme GitOps et en assurent le fonctionnement. Tandis que les conteneurs contiennent les applications logicielles, l'orchestrateur coordonne et assure le fonctionnement des conteneurs. Dans cette étape sont produits des artefacts tels que la spécification de la plateforme et des outils.

1. https://github.com/Azure/opendigitaltwins-dtdl
2. https://eclipse.dev/vorto/

Étape 3 : Cette étape est une étape d'instanciation et requiert la complétude des étapes précédentes. Les fichiers produits précédemment sont analysés et décomposés afin d'instancier un modèle de JN et un modèle de plateforme, resp. *DT Model* et *Platform Model* dans la Figure 2. Ensuite, au travers de transformations de modèle à modèle, nous instancions un modèle de GitOps puis les modèles d'*orchestrateur* et de *conteneur*.

Étape 4 : Dans cette étape, les artefacts finals sont générés par des transformations de modèle à texte. Ils permettent de configurer, décrire et spécifier le comportement attendu de l'orchestrateur (stratégie de redondance, adaptation de la charge) et des conteneurs (outils installés, configuration des outils).

Un prototype d'implémentation basé sur l'usine Fischertechnik est en cours. Cet exemple est centré sur l'opération de connexion (*connect*) du cycle de vie et met en œuvre des outils comme *Message Queueing Telemetry Transport* (MQTT) pour la communication de données, *Kubernetes* en tant qu'orchestrateur et *Docker* en tant qu'outil de conteneurisation. Les transformations de modèle à texte sont réalisées avec *Acceleo* et celles modèle à modèle grâce à *ATL* et à l'environnement de modélisation d'Eclipse (*EMF*). Une vue globale de ce processus est proposée en Figure 2.

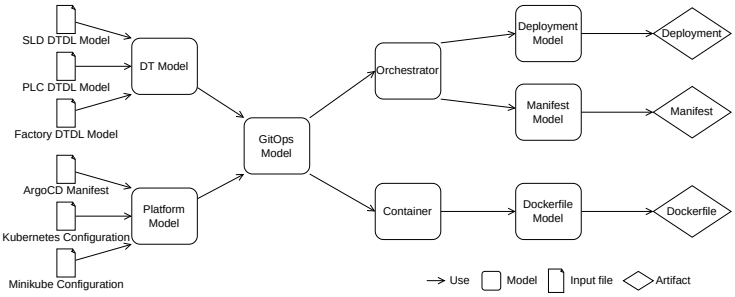


FIGURE 2. *Vue globale du processus d'automatisation du cycle de vie*

Ce papier propose un processus capable de prendre en compte le cycle de vie opérationnel d'un JN et d'en automatiser la gestion grâce à la génération d'artefacts logiciels pouvant être directement utilisés par des outils.

Bibliographie

Aissat S., Beaulieu J., Poirier E., Motamedi A., Gascon-Samson J., Bordeleau F., « A devops framework for the systematic engineering and evolution of digital twins for built assets », *Software and Systems Modeling*, October, 2025.

Beaumont G., Beugnard A., Martínez S., Urtado C., Vauttier S., « Towards Re-Engineering Digital Twins : Preliminary Experiments on Three Use Cases », *Proceedings of the ACM/IEEE 27th International Conference on Model Driven Engineering Languages and Systems*, MODELS Companion '24, Association for Computing Machinery, p. 453-458, October, 2024.

Quand les Métriques Dictent les Méthodes

Mitigation du biais de genre dans les systèmes NLP

Livia Leroy-Stone, Rébecca Deneckère

Centre de Recherche en Informatique
Université Paris 1 Panthéon-Sorbonne
livia.leroystone@gmail.com, rebecca.deneckere@univ-paris1.fr

RESUME. Les modèles de traitement automatique des langues sont largement intégrés dans les systèmes d'information et peuvent reproduire ou amplifier des biais de genre. De nombreuses méthodes de mitigation ont été proposées, mais leur diversité et l'hétérogénéité de leurs protocoles d'évaluation rendent leur comparaison et leur adoption difficiles dans des contextes opérationnels. Ce travail propose une analyse structurée des stratégies de mitigation du biais de genre en NLP, en mettant en évidence le rôle central joué par les métriques d'évaluation dans l'orientation et la conception des méthodes. À partir d'une revue systématique de la littérature, nous analysons les approches de pré- et de post-traitement selon leur stade d'intervention, les types de biais qu'elles ciblent effectivement et les métriques mobilisées pour en mesurer l'efficacité. L'analyse montre que les performances rapportées sont fortement dépendantes des métriques utilisées et qu'aucune méthode ne constitue une solution universelle. Nous discutons les implications de cette dépendance aux métriques pour la conception, l'intégration et la gouvernance des systèmes d'information intégrant des modèles de langage.

ABSTRACT. Natural language processing models are increasingly embedded in information systems, where they may reproduce or amplify gender bias. Numerous mitigation methods have been proposed, yet their diversity and the heterogeneity of evaluation protocols make comparison and adoption in operational settings challenging. This work provides a structured analysis of gender bias mitigation strategies in NLP, highlighting the central role of evaluation metrics in shaping and constraining methodological choices. Based on a systematic literature review, we examine pre-processing and post-processing approaches according to their intervention stage, the types of bias they actually address, and the metrics used to assess their effectiveness. The analysis shows that reported performance gains are strongly metric-dependent and that no mitigation strategy can be considered universal. We discuss the implications of this metric dependence for the design, integration, and governance of information systems relying on language models.

MOTS-CLES : NLP, Biais de genre, Revue systématique de littérature

KEYWORDS: NLP, Gender Bias, Systematic Literature Review

1. Introduction

Les modèles de traitement automatique des langues (NLP) sont aujourd'hui intégrés dans un nombre croissant de systèmes d'information (SI), où ils soutiennent des fonctions telles que la recherche d'information, l'aide à la décision, la recommandation, l'automatisation de traitements documentaires ou l'interaction homme-machine (Jorg *et al.*, 2023), (Ghassemi *et al.*, 2021). Leur adoption rapide, en particulier sous la forme de modèles pré-entraînés et de services largement diffusés, a renforcé leur rôle comme composants transverses de nombreux systèmes organisationnels (Gable, 2025), (OpenAI, 2023). De nombreux travaux ont montré que ces modèles peuvent reproduire, voire amplifier, des biais de genre présents dans les données d'entraînement, les ressources linguistiques ou les choix de conception des modèles (Bolukbasi *et al.*, 2016), (Zhao *et al.*, 2018), (Dev et Phillips, 2019), (Schick et Schütze, 2021). Lorsque ces modèles sont intégrés dans des SI opérationnels, ces biais ne relèvent plus uniquement d'un problème algorithmique, mais deviennent des enjeux systémiques susceptibles d'affecter la qualité des décisions, l'équité des services rendus et la confiance accordée aux systèmes (Gallegos *et al.*, 2024), (Raji *et al*., 2020). La mitigation du biais de genre constitue ainsi un problème à la fois technique et organisationnel, qui doit être analysé au regard des contraintes propres aux SI.

Il existe des méthodes visant à réduire le biais de genre dans les systèmes de traitement automatique des langues. Ces méthodes interviennent à différents stades du processus d'apprentissage, classiquement distingués en pré-traitement, in-processing et post-traitement (Friedler *et al.*, 2019) (Popović *et al.*, 2020) (Chen *et al.*, 2024) (Mehrabi *et al.*, 2021) (Li *et al.*, 2024) (Sobhani et Delany, 2024). Des revues de littérature existantes ont proposé des synthèses globales de ces approches (Sun *et al.*, 2019) ou se sont concentrées sur des catégories spécifiques, notamment les méthodes d'in-processing (Wan *et al.*, 2022) ou les techniques de détection du biais (Stanczak et Augenstein, 2021) (Gallegos *et al.*, 2024). Toutefois, ces travaux mettent également en évidence une forte hétérogénéité des méthodes proposées, tant dans leurs objectifs que dans leurs modalités d'évaluation. Les méthodes de mitigation ciblent des formes de biais différentes, qu'il s'agisse de biais directs ou indirects, statistiques ou causaux (Bolukbasi *et al.*, 2016) (Chen *et al.*, 2024) (Kumar *et al.*, 2020), et sont évaluées à l'aide de métriques variées, telles que WEAT, CrowS-Pairs, StereoSet ou WinoBias (Caliskan *et al.*, 2017) (Nadeem *et al.*, 2020) (Nangia *et al.*, 2020) (Zhao *et al*., 2018). Cependant ces métriques ne capturent qu'une partie des biais possibles et reposent sur des hypothèses implicites rarement explicitées. Cette dépendance aux métriques rend les résultats difficiles à comparer et peut masquer un décalage entre les objectifs normatifs affichés par les méthodes et les biais effectivement mesurés (Blodgett *et al.,* 2020).

Dans les systèmes d'information, ce décalage complique le choix raisonné d'une stratégie de mitigation. Les méthodes diffèrent non seulement par leur efficacité mesurée, mais aussi par leur coût de mise en œuvre, leur impact sur les performances globales, leur dépendance aux données d'entraînement ou leur facilité d'intégration dans des systèmes existants (Li *et al.*, 2024) (Ribeiro *et al.*, 2020). En particulier, les

approches de pré-traitement et de post-traitement présentent des profils contrastés en termes de contraintes de déploiement, ce qui les rend particulièrement pertinentes à analyser dans une perspective systèmes. Dans cet article, nous proposons une analyse structurée des méthodes de mitigation du biais de genre en traitement automatique des langues, en nous concentrant sur les approches de pré- et de post-traitement. À partir d'une revue systématique de la littérature, nous analysons ces méthodes selon trois axes complémentaires : le stade d'intervention dans le processus, les types de biais qu'elles permettent effectivement de traiter, et les métriques mobilisées pour évaluer leur efficacité. L'objectif n'est pas de proposer une nouvelle méthode, mais de fournir un cadre d'analyse permettant de comparer des approches hétérogènes et d'éclairer les choix de conception, d'intégration et d'exploitation de modèles de langage au sein de SI. Au-delà d'un état de l'art structuré, ce travail défend l'idée que les métriques d'évaluation jouent un rôle prescriptif dans la conception même des méthodes de mitigation. Loin d'être de simples outils de mesure, elles orientent les hypothèses retenues, les types de biais effectivement traités et, in fine, les compromis acceptés lors de l'intégration des modèles de langage dans des SI.

L'article est organisé comme suit. La section 2 présente les fondamentaux et la section 3 les travaux connexes. La section 4 décrit la méthodologie de recherche. La section 5 présente les résultats de l'analyse. Enfin, la section 6 propose une discussion et nous concluons dans la section 7.

2. Fondamentaux

Cette section présente les notions de biais considérées et les axes analytiques retenus pour comparer les approches étudiées.

Biais de genre en traitement automatique des langues. Dans la littérature, le biais de genre est généralement défini comme la tendance d'un système à favoriser ou défavoriser un genre par rapport à un autre (Moss-Racusin *et al.*, 2012). Dans le contexte NLP, cette notion recouvre des phénomènes variés, allant de déséquilibres statistiques dans les données d'entraînement à des associations stéréotypées entre termes genrés et concepts supposés neutres (Bolukbasi *et al.*, 2016) (Dev et Phillips, 2019). Les modèles de langage apprennent et reproduisent ces biais à partir de leurs données et de leurs représentations internes, notamment sous la forme de proximités sémantiques non désirées dans les espaces d'embedding (Bolukbasi *et al.*, 2016) (Kumar *et al.*, 2020). Lorsque ces modèles sont intégrés dans des systèmes d'information, ces biais peuvent influencer des décisions ou des interactions à grande échelle, ce qui confère à la mitigation du biais de genre une dimension à la fois technique et organisationnelle (Gallegos *et al.*, 2024).

Typologie des biais considérés. Afin de comparer les méthodes de mitigation de manière cohérente, nous distinguons plusieurs types de biais couramment mobilisés dans la littérature. Une première distinction oppose les biais directs aux biais indirects (Bolukbasi *et al.*, 2016) (Kumar *et al.*, 2020). Les biais directs correspondent à des associations explicites entre des termes genrés et des concepts censés être neutres, par exemple l'association entre « femme » et « cuisine ». Les biais indirects renvoient à

des regroupements sémantiques persistants entre concepts partageant une même connotation genrée, même en l'absence de marqueurs explicites. Par exemple, certaines professions peuvent être implicitement rapprochées d'un genre sans mention explicite. Une seconde distinction concerne les biais statistiques et les biais causaux (Chen *et al.*, 2024). Les biais statistiques sont mesurés à partir de différences observées dans les sorties des modèles entre groupes protégés. Les biais causaux visent à capturer des situations où la prédiction dépend directement d'un attribut protégé. Ces distinctions sont essentielles pour interpréter les résultats de mitigation, une réduction du biais statistique ne garantissant pas nécessairement l'absence de biais causal.

Stades d'intervention dans le processus d'apprentissage. Comme dit plus haut, les méthodes sont classées selon le stade du processus sur lequel elles interviennent : pré-traitement, in-processing et post-traitement (Friedler *et al.*, 2019) (Popović *et al.*, 2020) (Chen *et al.*, 2024) (Li *et al.*, 2024), (Sobhani et Delany, 2024). Le pré-traitement agit sur la collecte et la préparation des données d'entraînement, l'in-processing modifie le processus d'apprentissage, tandis que le post-traitement intervient lors de la configuration, de l'utilisation ou de l'exploitation des modèles. Dans ce travail, nous nous concentrons sur les approches de pré- et de post-traitement. Ces deux stratégies présentent des profils contrastés en termes de contraintes de déploiement et d'intégration dans des SI, en particulier lorsque les modèles sont pré-entraînés ou fournis sous forme de services externes. Ce contexte renforce l'intérêt des approches de post-traitement, particulièrement adaptées aux modèles pré-entraînés largement utilisés dans les systèmes actuels. La Figure 1 synthétise ces stades d'intervention ainsi que les contraintes associées et met en évidence que le choix d'un stade d'intervention conditionne directement les contraintes d'intégration, de maintenance et de gouvernance des modèles de langage dans les SI.

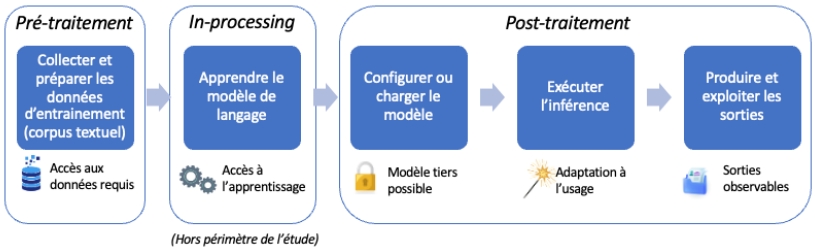


Figure 1 – Processus NLP et stades d'intervention des stratégies de mitigation du biais de genre.

Axes d'analyse retenus. Sur la base de ces éléments, l'analyse proposée repose sur trois axes complémentaires : le stade d'intervention dans le processus, les types de biais effectivement ciblés par les méthodes, et les métriques mobilisées pour évaluer leur efficacité. Ces axes structurent l'analyse comparative menée dans les sections suivantes en les plaçant dans une perspective de conception et d'exploitation de SI intégrant des modèles de langage.

3. Travaux connexes

Les travaux sur le biais de genre en traitement automatique des langues peuvent être regroupés en trois axes principaux : la détection et la mesure du biais, la mitigation du biais selon différents stades du processus d'apprentissage, et les analyses critiques des notions de biais et des métriques mobilisées dans la littérature.

Mise en évidence et quantification des biais de genre présents dans les représentations et les modèles de langage. Des travaux récents confirment également la persistance de ces biais dans les grands modèles de langage, y compris dans des contextes francophones (Bouchouchi *et al.,* 2026). Les biais dans les embeddings ont été étudiés à travers l'analyse de proximités sémantiques et d'analogies reflétant des stéréotypes (Bolukbasi *et al.*, 2016), (Dev et Phillips, 2019), ainsi que via des métriques d'association telles que WEAT et ses extensions aux encodeurs de phrases (Caliskan *et al.*, 2017) (May *et al.*, 2019). D'autres travaux ont proposé des benchmarks ciblant des manifestations plus contextuelles du biais, par exemple dans des tâches de résolution de coréférences (Zhao *et al.*, 2018). Des synthèses ont recensé ces approches de détection et d'évaluation, y compris dans le contexte des grands modèles de langage (Gallegos *et al.*, 2024), (Stanczak et Augenstein, 2021).

Mitigation du biais de genre. Une première revue de littérature a proposé une synthèse globale des méthodes de mitigation en NLP, couvrant différents stades du processus (Sun *et al.*, 2019). Des travaux plus récents se sont concentrés sur des catégories spécifiques de méthodes, en particulier l'in-processing, en analysant les modifications de l'apprentissage et les contraintes associées (Wan *et al.*, 2022). En parallèle, de nombreuses contributions proposent des méthodes ciblées intervenant sur les données, par exemple via augmentation ou substitution contrefactuelles (Zhao *et al.*, 2018), (Lu *et al.*, 2018), (Maudslay *et al.*, 2019), (Sobhani et Delany, 2024), sur les représentations, notamment par débiaisage des embeddings (Bolukbasi *et al.*, 2016), (Kumar *et al.*, 2020), (Wang *et al.*, 2020), ou à l'inférence, par des techniques de prompt tuning ou de neutralisation en test-time (Schick *et al.*, 2021), (Shen et Schütze, 2021), (Xie et Lukasiewicz, 2023).

Posture critique vis-à-vis des notions de biais et des métriques d'évaluation utilisées. Il a été montré que les définitions du biais et les objectifs normatifs associés sont hétérogènes et rarement explicités, ce qui complique la comparaison des méthodes et l'interprétation des résultats (Blodgett *et al.,* 2020). D'autres analyses soulignent que certaines méthodes peuvent réduire des scores de biais sans éliminer les structures sémantiques stéréotypées sous-jacentes (Gonen et Goldberg, 2019), ou que la diffusion large de benchmarks d'évaluation peut fragiliser la validité des résultats rapportés pour les modèles récents (Schick et Schütze, 2021).

Par rapport à ces travaux, notre contribution ne consiste pas à proposer une nouvelle méthode de mitigation, mais à analyser de manière structurée les approches de pré- et de post-traitement en mettant en regard le stade d'intervention, les types de biais effectivement ciblés et les métriques utilisées pour juger leur efficacité. L'objectif est de clarifier les conditions de comparabilité des résultats et d'éclairer les compromis associés aux choix de stratégies de mitigation.

4. Méthode de Recherche

Cette section décrit la méthodologie adoptée pour constituer le corpus d'articles analysés et structurer l'analyse des méthodes de mitigation du biais de genre en traitement automatique des langues. La démarche s'inspire des principes de transparence et de sélection progressive mobilisés dans les revues systématiques de type PRISMA, tout en étant adaptée au périmètre et aux objectifs de cette étude. La question de recherche principale de ce travail est la suivante: **Comment les méthodes de mitigation du biais de genre en traitement automatique des langues, en particulier celles fondées sur le pré- et le post-traitement, sont-elles structurées et évaluées dans la littérature, et quelles implications ces choix ont-ils pour leur intégration dans des systèmes d'information opérationnels ?**

Pour affiner l'analyse, nous avons décliné cette question en trois sous-questions de recherche : ***RQ1.*** Quelles stratégies de mitigation du biais de genre sont mobilisées dans la littérature selon le stade d'intervention dans le processus d'apprentissage ? ***RQ2.*** Quelles métriques et quels benchmarks sont utilisés pour évaluer l'efficacité des méthodes de mitigation, et quels types de biais permettent-ils effectivement de mesurer ? ***RQ3.*** Quelles sont les implications de ces choix méthodologiques et évaluatifs pour la conception, le déploiement et la gouvernance des systèmes d'information intégrant des modèles de langage ?

Les articles analysés ont été identifiés à partir de la base de données Scopus, retenue pour sa couverture large des principales revues et conférences en informatique, SI et intelligence artificielle. Ce choix visait à privilégier une base unique offrant une couverture large et homogène, afin de limiter les doublons et de conserver un processus de sélection cohérent. La recherche visait à identifier des travaux portant explicitement sur la mitigation du biais de genre dans les systèmes de traitement automatique des langues. La requête combinait, dans les titres, résumés et mots-clés, un ensemble de termes relatifs à la mitigation, à la réduction ou à la correction du biais, avec des termes renvoyant au genre, au sexisme, aux stéréotypes et aux discriminations, tandis que les termes liés au traitement automatique des langues, aux modèles de langage et aux embeddings étaient contraints au titre afin de garantir la pertinence thématique des résultats. La requête finale a permis d'identifier 190 articles. Bien que la requête ait été réalisée en février 2025, l'objectif étant une analyse structurée des approches plutôt qu'un recensement exhaustif, les tendances identifiées restent représentatives. La figure 2 illustre le processus et souligne le caractère progressif et restrictif de la sélection, visant à garantir la comparabilité des méthodes analysées plutôt qu'une couverture exhaustive de la littérature.

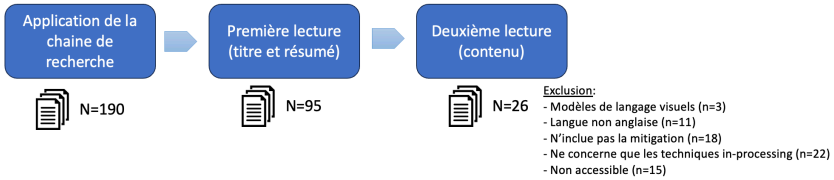


Figure 2 – Processus de sélection des articles analysés dans la revue systématique.

Une première étape de filtrage a consisté en une lecture des titres et résumés afin d'éliminer les travaux hors du périmètre NLP, ainsi que ceux portant uniquement sur la détection du biais ou relevant de revues de littérature générales. Les approches se concentrant exclusivement sur des modèles vision-langage ont également été exclues. La sélection a ensuite été affinée à partir de critères d'inclusion et d'exclusion portant sur le texte intégral des articles. Seuls les travaux proposant explicitement des méthodes de mitigation du biais de genre pour des modèles de langage textuels ont été retenus. L'analyse a été limitée aux articles portant sur la langue anglaise, afin de permettre une comparaison cohérente des méthodes et des métriques mobilisées.

Les travaux se concentrant exclusivement sur des méthodes d'in-processing ont été exclus, ces approches ayant déjà fait l'objet de synthèses dédiées (Wan *et al.*, 2022). Ce choix n'est pas uniquement motivé par des considérations de périmètre, mais par l'hypothèse que les approches de pré- et de post-traitement sont celles pour lesquelles la dépendance aux métriques est la plus structurante. Ces méthodes, fréquemment mobilisées dans des contextes où le contrôle sur les modèles est limité, illustrent de manière particulièrement nette la manière dont les métriques conditionnent les stratégies de mitigation adoptées. À l'issue de ce processus, 26 articles publiés entre 2016 et 2024 ont été retenus pour l'analyse. Ce corpus restreint reflète un compromis entre exhaustivité et comparabilité fine des méthodes analysées.

L'analyse du corpus est structurée autour des trois questions de recherche. Pour chaque article, une grille d'analyse a été appliquée afin d'extraire de manière systématique les informations pertinentes au regard des objectifs de l'étude. La grille d'analyse comprenait notamment le type de méthode, le stade d'intervention, les métriques utilisées, les types de biais ciblés et les contraintes d'intégration.

Ce travail s'inscrit dans une démarche analytique et comparative et ne vise pas à proposer une validation expérimentale des méthodes étudiées.

5. Analyse et Résultats

Cette section présente les résultats de l'analyse comparative des méthodes de mitigation du biais de genre identifiées dans le corpus. Elle s'organise autour des trois sous-questions de recherche portant respectivement sur les stratégies d'intervention mobilisées, les métriques utilisées pour évaluer leur efficacité et les implications de ces choix pour la conception et l'exploitation des systèmes d'information.

5.1. Stratégies de mitigation selon le stade d'intervention (RQ1)

Cette section analyse les méthodes de mitigation selon leur stade d'intervention.

Méthodes de pré-traitement. Elles reposent sur l'hypothèse selon laquelle une part importante du biais observé provient des déséquilibres et stéréotypes présents dans les données d'entraînement. Elles visent à modifier ces données avant l'apprentissage, par des mécanismes de transformation ou de rééquilibrage. Les approches *contrefactuelles*, telles que la Counterfactual Data Augmentation ou la substitution de noms genrés, constituent la famille la plus représentée dans le corpus (Zhao *et al.*,

2018) (Maudslay *et al.*, 2019) (Sobhani et Delany, 2024). Elles permettent de réduire efficacement certains biais mesurés par des benchmarks orientés tâches, notamment dans des contextes de génération ou de coréférence. Leur efficacité est toutefois étroitement liée aux règles de transformation appliquées et à la qualité des entités manipulées, ce qui limite leur généralisation et peut introduire des effets secondaires sur la cohérence linguistique. Les méthodes fondées sur le *masquage ou la neutralisation explicite des marqueurs de genre* cherchent à rendre les modèles moins sensibles aux informations genrées (Shen et Schütze, 2021), (Thakur *et al.*, 2023). Si elles permettent de réduire certains biais directs, leur impact reste partiel sur des métriques capturant des biais indirects ou contextuels. Les approches de rééquilibrage corrigent des déséquilibres de représentation sans modifier le contenu textuel, mais reposent sur une définition explicite des groupes et peinent à traiter des associations sémantiques plus complexes. Ces méthodes présentent l'avantage d'agir en amont et de produire des modèles intrinsèquement moins biaisés. Elles impliquent en revanche des coûts élevés en termes de préparation des données et de réentraînement, ce qui limite leur applicabilité dans des contextes où les modèles sont pré-entraînés ou fournis sous forme de services.

Méthodes de post-traitement. Elles interviennent après l'apprentissage du modèle et considèrent généralement celui-ci comme une boîte noire ou semi-transparente. Elles sont particulièrement pertinentes dans des contextes où l'accès aux données d'entraînement ou au processus d'apprentissage est restreint. Les approches de *débiaisage des représentations*, notamment des embeddings, visent à neutraliser un sous-espace de genre identifié dans l'espace vectoriel (Bolukbasi *et al.*, 2016) (Kumar *et al.*, 2020) (Wang *et al.*, 2020). Elles obtiennent des résultats probants sur des tests d'association comme WEAT, mais leur impact sur des tâches aval et sur des biais indirects reste limité et dépend fortement de la définition du sous-espace et des termes considérés comme neutres. Les méthodes explicitement guidées par une métrique illustrent de manière encore plus marquée la dépendance aux outils d'évaluation, en optimisant directement les scores mesurés sans garantie de réduction plus générale du biais (Popović *et al.*, 2020) (Gonen et Goldberg, 2019). Les techniques de *post-traitement à l'inférence*, telles que le prompt tuning ou les mécanismes de neutralisation en test-time, offrent une grande flexibilité et un coût de déploiement réduit (Schick *et al.*, 2021) (Shen et Schütze, 2021) (Xie et Lukasiewicz, 2023). Elles permettent d'agir sur des scénarios d'usage spécifiques, en particulier dans des systèmes intégrant des modèles tiers, mais leur effet reste souvent contextuel et dépendant des métriques retenues pour l'évaluation. Les méthodes de post-traitement offrent un compromis favorable en termes de déploiement et d'intégration dans des SI, au prix d'une efficacité plus ciblée et d'une forte dépendance aux conditions d'évaluation.

Comparaison transversale. L'analyse des méthodes de pré- et de post-traitement met en évidence qu'elles ciblent des formes de biais différentes et reposent sur des hypothèses distinctes quant à l'origine du biais. Les méthodes de pré-traitement cherchent à corriger des déséquilibres à la source, tandis que les méthodes de post-traitement visent à ajuster les représentations ou les sorties sans remettre en cause le modèle sous-jacent. Aucune approche ne permet de traiter simultanément l'ensemble des formes de biais identifiées, et les performances rapportées dépendent fortement des

métriques utilisées pour l'évaluation. Cette observation souligne la nécessité d'une lecture critique des résultats de la littérature et justifie l'analyse des métriques et de leurs limites, abordée dans la section suivante. Ces différences ne relèvent pas uniquement de choix techniques. Elles conditionnent la capacité à auditer les modèles, à maintenir les stratégies de mitigation dans le temps et à aligner les performances mesurées avec les objectifs organisationnels du système.

5.2. Métriques d'évaluation et types de biais mesurés (RQ2)

L'évaluation des méthodes de mitigation repose sur un ensemble hétérogène de métriques qui ne mesurent pas les mêmes formes de biais et participent à leur définition opérationnelle. Le tableau 1 offre une analyse croisée entre les méthodes de mitigation et les métriques d'évaluation. Cette spécialisation suggère que de nombreuses approches sont implicitement conçues pour optimiser des scores spécifiques, au risque de négliger des formes de biais non couvertes par les benchmarks retenus. La notion de couverture désigne la capacité d'une méthode à améliorer les scores sur une métrique donnée, telle que rapportée dans la littérature. Cette catégorisation repose sur une synthèse qualitative des résultats et ne constitue pas une mesure quantitative standardisée.

Tableau 1. Couverture relative des méthodes selon les métriques d'évaluation.

Méthodes \ Métriques	WEAT	CrowS-Pairs	StereoSet	WinoBias
Pré-traitement				
CDA / CDS	◐	●	●	●
Masquage	◐	◐	◐	○
Rééquilibrage	◐	◐	◐	◐
Post-traitement				
Débiaisage embeddings	●	◐	◐	○
Optimisation guidée par WEAT	●	○	○	○
Inférence / prompt	○	●	●	◐

Couverture : ● forte (amélioration significative rapportée) / ◐ partielle (amélioration partielle ou dépendante du contexte) / ○ faible ou non ciblée (peu ou pas d'effet observé)
Type de Biais : Bleu = biais direct / Orange = biais indirect / Violet = biais causal - décisionnel

Tests d'association dans les espaces de représentation. Les tests d'association, tels que le Word Embedding Association Test (WEAT) et ses extensions aux encodeurs de phrases (SEAT), mesurent les associations entre des ensembles de termes genrés et des concepts supposés neutres à partir de distances dans l'espace vectoriel (Caliskan *et al.*, 2017) (May *et al.*, 2019). Ces métriques sont principalement utilisées pour évaluer des méthodes agissant sur les représentations, notamment le débiaisage des embeddings (Bolukbasi *et al.*, 2016) (Kumar *et al.*, 2020) (Popović *et al.*, 2020). Si ces tests permettent de détecter des biais directs et statistiques dans les représentations, leur portée reste limitée à des associations prédéfinies et ne capture ni les biais indirects plus complexes ni les effets contextuels. Plusieurs travaux ont en outre montré que l'optimisation directe de ces métriques peut conduire à une réduction

artificielle des scores de biais sans suppression des structures sémantiques stéréotypées sous-jacentes (Gonen et Goldberg, 2019).

Benchmarks fondés sur des paires de phrases stéréotypées. D'autres métriques reposent sur des jeux de données constitués de paires de phrases stéréotypées et anti-stéréotypées, telles que CrowS-Pairs ou StereoSet (Nadeem *et al.*, 2020) (Nangia *et al.*, 2020). Ces benchmarks évaluent la préférence d'un modèle pour des formulations stéréotypées dans des contextes proches de scénarios d'usage réels, en particulier pour des tâches de génération ou de complétion. Ces métriques permettent de capturer des biais directs et indirects dans les sorties des modèles et sont fréquemment mobilisées pour évaluer des méthodes de pré-traitement et de post-traitement à l'inférence (Li *et al.*, 2024), (Schick *et al.*, 2021). Leur principal avantage réside dans leur proximité avec les usages applicatifs, mais leur validité dépend fortement de la couverture des jeux de données et de leur indépendance vis-à-vis des corpus d'entraînement des modèles récents, ce qui peut fragiliser l'interprétation des résultats (Schick et Schütze, 2021).

Tâches de résolution de coréférences. Certaines métriques évaluent le biais à travers des tâches spécifiques, telles que la résolution de coréférences avec le benchmark WinoBias (Zhao *et al.*, 2018). Ces approches permettent de mettre en évidence des formes de biais indirects et causaux, en mesurant l'impact du genre sur des décisions prises par le modèle dans des contextes ambigus. Bien que ces métriques soient particulièrement pertinentes pour analyser des effets décisionnels, leur champ d'application reste limité à une tâche donnée, ce qui complique la généralisation des résultats. Les gains observés sur ces benchmarks ne garantissent pas une réduction du biais dans d'autres contextes d'usage ou pour d'autres types de décisions.

Impact des métriques sur l'interprétation des résultats. L'analyse transversale des métriques utilisées dans la littérature montre que l'efficacité d'une méthode de mitigation est étroitement dépendante des outils d'évaluation retenus. De nombreuses approches obtiennent de bons résultats sur les métriques qu'elles ciblent explicitement, sans que ces gains se traduisent nécessairement par une réduction plus générale du biais (Blodgett *et al.,* 2020). Cette dépendance pose un problème de gouvernance des modèles. Le choix d'une métrique revient souvent à adopter implicitement une définition particulière du biais, qui peut ne pas être alignée avec les objectifs fonctionnels, organisationnels ou réglementaires du système concerné. Dans des contextes à fort impact décisionnel, une interprétation non critique des scores de biais peut conduire à une fausse impression de neutralité, alors que certaines formes de biais persistent ou se déplacent.

Ces constats soulignent la nécessité d'une lecture prudente des résultats de mitigation et d'une explicitation des hypothèses sous-jacentes aux métriques utilisées, en particulier lorsque les modèles de langage sont intégrés dans des systèmes d'information opérationnels. Cette dépendance aux métriques ne pose pas seulement un problème d'interprétation des résultats expérimentaux. Lorsqu'ils sont intégrés dans des SI opérationnels, ces choix évaluatifs ont des conséquences directes sur la gouvernance des modèles, la confiance accordée aux systèmes et la nature des décisions qu'ils soutiennent, ce qui motive l'analyse développée dans la section suivante.

5.3. Implications pour les systèmes d'information (RQ3)

Les analyses précédentes montrent que la mitigation du biais de genre dans les systèmes de traitement automatique des langues ne peut être abordée comme un simple problème d'optimisation algorithmique. Lorsqu'ils sont intégrés dans des SI opérationnels, les modèles de langage deviennent des composants sociotechniques dont les biais ont des effets à la fois techniques, organisationnels et décisionnels (Gallegos *et al.*, 2024). Cette section discute les implications de ces constats pour la conception, le déploiement et la maintenance de SI intégrant des modèles NLP. Les principaux compromis associés aux stratégies de pré- et de post-traitement sont synthétisés dans la Figure 4. En particulier, la couverture plus limitée des biais par les stratégies de post-traitement découle des observations présentées dans les sections précédentes, notamment de leur dépendance aux métriques retenues et aux scénarios d'usage.

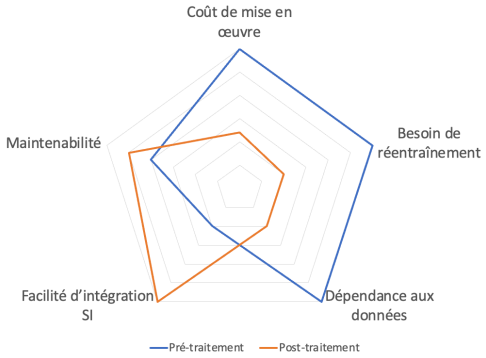


Figure 4 – Profils de compromis des stratégies de mitigation du biais de genre.

Cette figure met en évidence que les stratégies de post-traitement privilégient la flexibilité au détriment de la couverture des biais, tandis que les approches de pré-traitement offrent une meilleure maîtrise au prix de coûts organisationnels plus élevés.

Choix des stratégies de mitigation et contraintes de déploiement. Le choix d'une stratégie de mitigation dépend du degré de contrôle dont disposent les concepteurs sur les données et les modèles. Les méthodes de pré-traitement supposent un accès aux données d'entraînement et impliquent souvent un réentraînement du modèle (Zhao *et al.*, 2018), (Lu *et al.*, 2018). Elles sont donc plus adaptées à des contextes où les modèles sont développés en interne ou finement adaptés à un domaine spécifique. Ces approches peuvent s'avérer coûteuses en termes de calcul et de préparation des données et difficiles à maintenir dans le temps lorsque les corpus évoluent. À l'inverse, les méthodes de post-traitement sont particulièrement attractives dans des contextes où les modèles sont pré-entraînés ou fournis sous forme de services externes, ce qui est fréquent dans les SI contemporains (Li *et al.*, 2024). Leur déploiement est généralement plus simple et moins coûteux, mais leur efficacité dépend fortement des scénarios d'usage et des métriques retenues pour l'évaluation. Ce compromis entre contrôle, coût et flexibilité doit être

explicitement pris en compte lors de la conception des systèmes. Par exemple, dans un système d'aide au recrutement ou de recommandation de profils, une réduction du biais mesurée par une métrique comme WEAT peut coexister avec des décisions finales toujours influencées par des stéréotypes de genre. Ce type de décalage illustre les limites d'une approche exclusivement métrique dans des SI à fort impact décisionnel.

Dépendance aux métriques et gouvernance des modèles. L'analyse des métriques montre que les résultats de mitigation sont étroitement liés aux outils d'évaluation utilisés. Cela pose des questions de gouvernance des modèles et des données. Optimiser un système selon une métrique donnée revient à adopter implicitement une définition particulière du biais, qui peut ne pas correspondre aux objectifs organisationnels ou réglementaires du système (Blodgett *et al.,* 2020). Dans des contextes à fort impact décisionnel, tels que le recrutement, l'orientation ou la recommandation de contenus, cela peut conduire à une fausse impression de neutralité, alors que certaines formes de biais persistent ou se déplacent. La gouvernance des systèmes intégrant des modèles NLP devrait ainsi inclure une réflexion explicite sur les métriques utilisées, leur champ de validité et leurs limites, plutôt qu'une simple adoption de benchmarks standard.

Maintenabilité et dette technique liée à la mitigation du biais. Les stratégies de mitigation du biais ont également des implications en termes de maintenabilité des systèmes. Les méthodes de pré-traitement et d'in-processing tendent à produire des modèles spécialisés, dont la mise à jour nécessite de répéter des étapes coûteuses de préparation des données et d'apprentissage. À l'inverse, certaines méthodes de post-traitement, notamment celles fondées sur le prompt tuning ou la neutralisation en test-time, introduisent une couche supplémentaire de logique qui peut accroître la complexité du système et générer une dette technique spécifique (Schick *et al.*, 2021), (Shen et Schütze, 2021), (Xie et Lukasiewicz, 2023). Dans les deux cas, la mitigation du biais ne peut être considérée comme une opération ponctuelle. L'évolution des données, des usages et des attentes sociétales implique une réévaluation régulière des choix effectués. Cette dimension dynamique est rarement abordée dans la littérature technique, mais elle est centrale du point de vue de l'exploitation des systèmes d'information.

Vers une aide à la décision pour les concepteurs de systèmes. Les résultats présentés dans cet article suggèrent qu'il n'existe pas de stratégie de mitigation universelle applicable à tous les systèmes et à tous les contextes. Le choix d'une approche relève plutôt d'un arbitrage entre plusieurs dimensions : types de biais jugés critiques, contraintes de déploiement, coûts de maintenance, impact sur la performance et exigences organisationnelles. Le cadre d'analyse proposé vise à fournir une aide à la décision pour les concepteurs et responsables de SI intégrant des modèles de langage. En mettant en relation les stades d'intervention, les types de biais ciblés et les métriques mobilisées, il permet d'expliciter les compromis sous-jacents aux différentes stratégies de mitigation et d'éviter une adoption purement opportuniste ou métrique-dépendante des méthodes proposées dans la littérature.

6. Discussion

L'analyse met en évidence plusieurs enseignements transversaux. En particulier, elle montre que la notion même de biais, telle qu'elle est mobilisée dans la littérature,

recouvre des réalités hétérogènes, rarement explicitées de manière uniforme (Blodgett *et al.,* 2020). Cette hétérogénéité se reflète à la fois dans les objectifs assignés aux méthodes, dans les types de biais qu'elles ciblent effectivement et dans les métriques utilisées pour en évaluer l'efficacité.

Un premier point saillant concerne l'absence de correspondance systématique entre les objectifs normatifs affichés par les méthodes et les biais mesurés empiriquement. De nombreuses approches revendiquent une réduction des stéréotypes ou une amélioration de l'équité, tout en s'appuyant sur des métriques qui capturent principalement des associations statistiques ou des effets locaux dans des tâches spécifiques (Caliskan *et al.*, 2017) (Nadeem *et al.*, 2020) (Nangia *et al.*, 2020). Cette situation conduit à des résultats difficiles à comparer et peut donner une impression trompeuse de progrès, lorsque les gains observés sont fortement dépendants des benchmarks utilisés (Gonen et Goldberg, 2019).

Un second point concerne la complémentarité, mais aussi la tension, entre les méthodes de pré-traitement et de post-traitement. Les premières agissent en amont et peuvent produire des modèles intrinsèquement moins biaisés, mais au prix de coûts élevés en termes de données, de calcul et de maintenance (Zhao et al., 2018), (Lu et al., 2018). Les secondes offrent une plus grande flexibilité et une intégration facilitée dans des systèmes existants, mais leur efficacité reste souvent contextuelle et dépendante des scénarios d'usage (Schick et al., 2021), (Shen et Schütze, 2021), (Xie et Lukasiewicz, 2023). Cette opposition reflète des choix architecturaux et organisationnels qui relèvent davantage de la conception des SI que de l'optimisation algorithmique.

L'analyse met également en évidence un effet de spécialisation croissante des méthodes, en particulier dans les travaux récents. De nombreuses approches sont conçues pour optimiser un benchmark ou une métrique précise, ce qui améliore les performances mesurées mais limite la généralisation des résultats (Li *et al.*, 2024). Cette tendance est accentuée par la diffusion large de certains jeux de données d'évaluation, susceptibles d'être intégrés aux corpus d'entraînement des modèles récents, rendant l'interprétation des scores plus délicate (Schick et Schütze, 2021).

L'évolution des modèles, des données et des usages implique une réévaluation continue des choix effectués. Or cette dimension dynamique est rarement abordée dans la littérature technique, alors qu'elle constitue un enjeu central pour l'exploitation de SI à long terme. Ces résultats invitent à dépasser une approche purement métrique de la mitigation du biais. Dans des SI à fort impact décisionnel, l'optimisation de benchmarks standard peut produire une illusion de neutralité, alors même que certaines formes de biais persistent ou se déplacent hors du champ de mesure.

Enfin, la diversité des équipes impliquées dans la conception et l'évaluation des systèmes apparaît comme un facteur important pour identifier et traiter les biais en amont, bien que cet aspect reste encore peu abordé dans les travaux techniques.

Cette analyse se limite volontairement aux modèles et ressources en langue anglaise, ce qui reflète l'état actuel de la littérature dominante mais constitue une limite pour la généralisation des résultats à d'autres contextes linguistiques et culturels.

7. Conclusion

Ce travail propose une analyse structurée des méthodes de mitigation du biais de genre en NLP. À partir d'une revue systématique de la littérature, les méthodes ont été analysées selon trois axes complémentaires : le stade d'intervention dans le processus d'apprentissage, les types de biais effectivement ciblés et les métriques utilisées pour évaluer leur efficacité. L'analyse montre qu'aucune méthode ne s'impose comme une solution universelle et que les performances rapportées dépendent fortement de l'alignement entre stratégie de mitigation, définition du biais et outils d'évaluation. Ce constat montre la nécessité d'une lecture critique des résultats de la littérature et d'une explicitation des compromis sous-jacents, en particulier lorsqu'elles sont intégrées dans des systèmes d'information opérationnels. Les résultats soulignent également l'importance de considérer la mitigation du biais comme un problème de conception et de gouvernance des systèmes, et non comme un simple ajustement technique. Le cadre d'analyse proposé vise à aider les concepteurs et décideurs à sélectionner des stratégies de mitigation adaptées à leurs contextes d'usage, en tenant compte des contraintes de déploiement, de maintenance et d'évolution des systèmes. Dans des systèmes d'information à fort impact décisionnel, l'optimisation de benchmarks standard peut produire une illusion de neutralité, alors même que certaines formes de biais persistent ou se déplacent hors du champ de mesure.

L'analyse pourrait être étendue à des méthodes d'in-processing et à des approches combinées, afin d'évaluer si les tendances observées se confirment lorsque l'ensemble du processus est pris en compte (Wan *et al.*, 2022). L'étude de langues autres que l'anglais, ainsi que de définitions du genre dépassant la binarité, constitue un axe important pour améliorer la portée des méthodes existantes. Le développement de cadres d'évaluation plus explicitement alignés avec des objectifs organisationnels et sociétaux reste aussi un enjeu central pour une intégration responsable des modèles de langage dans les systèmes d'information.

Bibliographie

Blodgett S. L., Barocas S., Daumé III H., Wallach H. (2020). Language (technology) is power: A critical survey of "bias" in NLP ; Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL).

Bolukbasi T., Chang K.-W., Zou J., Saligrama V., Kalai A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings ; Advances in Neural Information Processing Systems (NeurIPS).

Bouchouchi N., Laugel T., Lesot M.-J., Marsala C., Renard X. (2026). Biais de genre encodés et exprimés dans les LLM : étude conjointe ; Extraction et Gestion des Connaissances (EGC), Anglet, France.

Caliskan A., Bryson J. J., Narayanan A. (2017). Semantics derived automatically from language corpora contain human-like biases ; Science, vol. 356, n° 6334, p. 183–186.

Chen H., Ji Y., Evans D. (2024). Addressing both statistical and causal gender fairness in NLP models ; Artificial Intelligence.

Dev S., Phillips J. M. (2019). Attenuating bias in word vectors ; 22nd International Conference on Artificial Intelligence and Statistics (AISTATS).

Friedler S. A., Scheidegger C., Venkatasubramanian S., Choudhary S., Hamilton E. P., Roth D. (2019). A comparative study of fairness-enhancing interventions in machine learning. Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (FAccT).

Gallegos I. O., Rossi R. A., Barrow J., et al. (2024). Bias and fairness in large language models: A survey ; Computational Linguistics, vol. 50, n° 3, p. 1097–1179.

Gonen H., Goldberg Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them ; arXiv preprint.

Gable P. (2025). L'application Le Chat de Mistral AI passe un cap symbolique avec un million de téléchargements. Article en ligne, https://www.lemonde.fr/ (consulté en avril 2025).

Ghassemi M., Oakden-Rayner L., Beam A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care ; The Lancet Digital Health, vol. 3, n° 11

Jörg T., Kämpgen B., Feiler D., Müller L., Düber C., Mildenberger P., Jungmann F. (2023). Efficient structured reporting in radiology using an intelligent dialogue system based on speech recognition and natural language processing ; Insights into Imaging, vol. 14, n° 1.

Ribeiro M. T., Wu T., Guestrin C., Singh S. (2020). Beyond accuracy: Behavioral testing of NLP models with CheckList ; Transactions of the Association for Computational Linguistics (TACL), vol. 8, p. 490–506.

Kumar V., Bhotia T. S., Kumar V., Chakraborty T. (2020). Nurse is closer to woman than surgeon? Mitigating gender-biased proximities in word embeddings ; 58th Annual Meeting of the Association for Computational Linguistics (ACL).

Li Y., Du M., Song R., Wang X., Sun M., Wang Y. (2024). Mitigating social biases of pre-trained language models via contrastive self-debiasing with double data augmentation ; Artificial Intelligence.

Lu K., Mardziel P., Wu F., Amancharla P., Datta A. (2018). Gender bias in neural natural language processing ; arXiv preprint.

Maudslay R. H., Gonen H., Cotterell R., Teufel S. (2019). It's all in the name: Mitigating gender bias with name-based counterfactual data substitution ; 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).

May C., Wang A., Bordia S., Bowman S. R., Rudinger R. (2019). On measuring social bias in sentence encoders ; 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT).

Mehrabi N., Morstatter F., Saxena N., Lerman K., Galstyan A. (2021). A survey on bias and fairness in machine learning ; ACM Computing Surveys, vol. 54, n° 6, article 115.

Moss-Racusin C. A., Dovidio J. F., Brescoll V. L., Graham M. J., Handelsman J. (2012). Science faculty's subtle gender biases favor male students ; National Academy of Sciences of the United States of America, vol. 109, n° 41.

Nadeem M., Bethke A., Reddy S. (2020). StereoSet: Measuring stereotypical bias in pretrained language models ; arXiv preprint.

Nangia N., Vania C., Bhalerao R., Bowman S. R. (2020). CrowS-Pairs: A challenge dataset for measuring social biases in masked language models ; 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).

OpenAI (2023). ChatGPT reaches one million users in two months. OpenAI Blog, https://openai.com/blog/ (consulté en 2025).

Popović R., Lemmerich F., Strohmaier M. (2020). Joint multiclass debiasing of word embeddings ; 2020 Conference on Fairness, Accountability, and Transparency (FAccT).

Raji I. D., Smart A., White R. N., et al. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing ; Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAccT).

Stanczak K., Augenstein I. (2021). A survey on gender bias in natural language processing ; 59th Annual Meeting of the Association for Computational Linguistics (ACL).

Schick T., Udupa S., Schütze H. (2021). Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP ; Transactions of the Association for Computational Linguistics (TACL).

Schick T., Schütze H. (2021). It's not just size that matters: Small language models are also few-shot learners ; 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT).

Shen T., Li J., Bouadjenek M. R., Mai Z., Sanner S. (2023). Towards understanding and mitigating unintended biases in language model-driven conversational recommendation ; Information Processing & Management.

Sobhani N., Delany S. (2024). Towards fairer NLP models: Handling gender bias in classification tasks ; 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP).

Sun T., Gaut A., Tang S., et al. (2019). Mitigating gender bias in natural language processing: Literature review ; 57th Annual Meeting of the Association for Computational Linguistics (ACL).

Thakur H., Jain A., Vaddamanu P., Liang P. P., Morency L.-P. (2023). Language models get a gender makeover: Mitigating gender bias with few-shot data interventions. 61st Annual Meeting of the Association for Computational Linguistics (ACL).

Wang T., Lin X. V., Rajani N. F., McCann B., Ordonez V., Xiong C. (2020). Double-hard debias: Tailoring word embeddings for gender bias mitigation ; 58th Annual Meeting of the Association for Computational Linguistics (ACL).

Wan M., Zha D., Liu N., Zou N. (2022). In-processing modeling techniques for machine learning fairness: A survey ; ACM Transactions on Knowledge Discovery from Data.

Xie Z., Lukasiewicz T. (2023). An empirical analysis of parameter-efficient methods for debiasing pre-trained language models ; arXiv preprint.

Zhao J., Wang T., Yatskar M., Ordonez V., Chang K.-W. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods ; 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT).

Revue systématique des critères d'évaluation des scénarios prospectifs

Brieuc Danet[1], Anouck Chan[1], Thomas Polacsek[1]

1. DTIS, ONERA, Université de Toulouse
2 avenue Marc Pellegrin, BP 74025 - 31055 Toulouse CEDEX 4 France
prénom.nom@onera.fr

RESUME.
Contexte *La* méthode des scénarios *consiste à confronter un utilisateur à un ensemble de scénarios prospectifs pour l'aider à appréhender les incertitudes de l'avenir et à prendre les décisions adéquates. L'élaboration et le choix de scénarios à prendre en compte répondent aux besoins que cet utilisateur exprime à travers un ensemble de critères de sélection.*
Objectif *Cette revue a pour objectif de recenser les critères utilisés pour évaluer les scénarios prospectifs, que ce soit pour leur élaboration ou pour leur sélection en vue d'être exploités.*
Méthodologie *Le protocole de revue a suivi les prérogatives de la méthode PRISMA.*
Résultats *Une liste de 104 critères a été établie parmi les 39 articles du corpus sélectionné.*
Conclusion *La revue a permis d'établir un éventail exhaustif des critères d'évaluation des scénarios prospectifs. Les travaux futurs devront permettre de déterminer, au sein des scénarios, les éléments clés sur lesquels reposent les critères.*

ABSTRACT.
Context Scenario planning *involves presenting a user with a set of prospective scenarios to help them understand future uncertainties and make appropriate decisions. The development and selection of scenarios address the needs expressed by this user through a set of selection criteria.*
Goal *This review aims to identify the criteria used to evaluate prospective scenarios, whether for their development or their selection for use.*
Methodology *The review protocol followed the prerogatives of the PRISMA method.*
Results *A list of 104 criteria was established among the 39 articles in the selected corpus.*
Conclusion *The review established a comprehensive set of criteria for evaluating scenarios. Future work should determine, within scenarios, the key elements on which the criteria rest.*

MOTS-CLÉS : *scénario, prospective, revue systématique, critère, modèle conceptuel*

KEYWORDS: *scenario, prospective, systematic review, criteria, conceptual model*

1. Introduction

1.1. *Contexte*

L'attitude prospective, initiée par le philosophe Gaston Berger dans les années 1960, consiste à *se préparer à l'action,* [en s'appliquant à] *dessiner d'une manière aussi rationnelle que possible les divers visages que pourrait prendre le monde de demain* (Berger, 1964). Elle a conduit à l'adoption d'une démarche visant à fournir aux organisations les visions de l'avenir leur permettant de prendre les meilleures décisions face aux incertitudes dans leur domaine stratégique.
Ainsi est né le concept de scénario[1], en tant que *description cohérente de futur hypothétique alternatif servant de base à l'action* (Van Notten *et al.*, 2003), dont l'élaboration a donné lieu au développement de différentes méthodes adaptées aux besoins. Deux principales écoles ont mené les premières réflexions théoriques et applicatives pour tenter de formaliser et d'outiller ces méthodes des scénarios : aux Etats-Unis à la RAND[2] Corporation avec Herman Kahn, physicien et mathématicien, fondateur de l'Institut Hudson ; en France par l'OTAM[3] à l'initiative de la DATAR[4] (Julien *et al.*, 1975).

Initialement mise en œuvre pour la défense et l'aménagement du territoire, la démarche prospective s'est ensuite largement étendue à de nombreux autres secteurs : santé, éducation, économie, transport, environnement. Elle s'est appuyée sur le *scenario planning*, fondé dès les années 1990 sur la volonté d'*intégrer les scénarios de manière explicite aux différents processus de planification stratégique* (Schoemaker, 1991). Certaines organisations s'engagent ainsi dans la construction de scénarios dits *normatifs* (Tuominen *et al.*, 2014), dans le but de comprendre quelles décisions prendre pour atteindre un avenir souhaitable, comme l'OACI [5] pour établir des objectifs communs aux acteurs du transport aérien en vue de respecter les accords de Paris à l'horizon 2050 (ICAO/CAEP, 2022). D'autres préfèrent envisager des scénarios exploratoires pour mieux appréhender l'ensemble des avenirs possibles et ainsi *ne plus être surpris par la surprise* (Burt, 2007), une ambition affichée par le GIEC [6] pour évaluer l'impact à long terme de l'activité humaine sur le dérèglement climatique (Girod *et al.*, 2009).

La production de scénarios prospectifs s'est accélérée ces vingt dernières années dans le transport aérien suite à la prise de conscience de la vulnérabilité du secteur aux perturbations dans son environnement direct : tensions politiques autour de l'approvisionnement en matières premières, crise sanitaire du COVID. L'existence de ces scénarios devrait faciliter le travail d'acteurs souhaitant baser leur réflexion sur des visions futures, que ce soit par exemple les climatologues qui souhaitent intégrer des projections du trafic dans les modèles climatiques pour en évaluer l'impact, ou

1. De nombreuses acceptions du terme ont été proposées, de même que plusieurs classifications, sans qu'aucun standard n'ait pu être instauré (Cordova-Pozo and Rouwette, 2023)
2. Research ANd Development
3. Omnium Technique d'AMénagement
4. Délégation à l'Aménagement du Territoire et à l'Action Régionale
5. Organisation de l'Aviation Civile Internationale
6. Groupe d'Experts Intergouvernemental sur l'Evolution du Climat

les organismes de régulations pour identifier et planifier les leviers de réduction des émissions.

Pourtant, l'exploitation de ces scénarios reste aujourd'hui difficile en raison de leur hétérogénéité structurelle et formelle. Comment sélectionner les scénarios d'intérêt parmi un ensemble de propositions générées par des acteurs aux objectifs, aux méthodes et aux intérêts divergents ? Comment comparer des scénarios aux structures internes et aux hypothèses disparates (horizon considéré, variables prises en compte, périmètre couvert, etc) ?

1.2. *Motivation*

Que ce soit pour élaborer de nouveaux scénarios pertinents ou pour les sélectionner parmi les scénarios existants, des critères ont émergé, qui pointent des caractéristiques des scénarios devant permettre d'en juger la valeur au regard de leur intérêt présumé pour l'utilisateur. Certains critères, notamment liés à l'occurrence future des hypothèses envisagées, ont rapidement donné lieu à des méthodes spécifiques d'évaluation et d'élaboration. Parmi les premières en date, la méthode du *Cône des plausibles* développée par W. Taylor à l'US Army War College (Taylor, 1993) doit permettre d'assurer la plausibilité des scénarios créés en traçant les liens de cause à effet entre les événements impliqués dans les tendances futures.

Néanmoins, la grande variété des besoins a conduit à utiliser un nombre croissant de critères pour satisfaire plus spécifiquement chaque utilisateur dans le contexte de son étude prospective propre. La multiplication des critères utilisés pour juger de la qualité des scénarios soulève alors une question pratique fondamentale : selon quelles méthodes et sur la base de quels éléments sélectionner les scénarios les plus adaptés aux besoins de l'étude ? La difficulté est accrue par la nature essentiellement narrative des scénarios : comment en extraire les éléments pertinents pour les évaluer et les comparer entre eux, quelle qu'en soit la source ? Un formalisme jouant le rôle de grille d'analyse serait à cet égard utile. Il permettrait de représenter tout scénario sous une forme structurée, indépendante de sa forme originale. Un tel scénario formalisé pourrait alors être évalué selon les différents critères et grâce aux méthodes disponibles dans la littérature, et comparé à des scénarios issus de sources distinctes.

1.3. *Questions de recherche*

Pour définir un tel formalisme, il convient tout d'abord de recenser les critères d'évaluation existants afin d'identifier les éléments sur lesquels repose leur application. C'est sur la base de ces éléments qu'il sera possible par la suite d'élaborer un modèle conceptuel de scénario. Cette revue propose donc, d'une part de dresser une liste exhaustive des critères utilisés dans les études prospectives pour évaluer les scénarios, d'autre part d'en collecter les définitions afin d'expliciter les notions qu'ils recouvrent et les éléments qu'ils prennent en compte.

Nous avons ainsi mené une revue systématique pour établir un corpus complet de références bibliographiques permettant de répondre de façon exhaustive à ces objectifs, à travers deux questions :
Question Q1 : Quels sont tous les critères d'évaluation utilisés dans les méthodes d'élaboration et de sélection des scénarios futurs ?
Question Q2 : Quelles sont les définitions associées à chacun de ces critères ?

Le paragraphe 2 détaille les méthodes et processus employés pour rechercher et sélectionner les publications d'intérêt et en extraire une liste exhaustive de critères et de leurs définitions. Le paragraphe 3 présente les résultats obtenus. Le paragraphe 4 conclut sur les réponses apportées aux questions de recherche, discute des limites de la revue, et propose des perspectives à envisager pour l'analyse de la base de données obtenue.

2. Méthode

Le protocole adopté pour cette revue systématique s'appuie sur les recommandations PRISMA[7] (Page, 2021). Il s'organise en trois étapes : l'identification des publications, la sélection des publications, et l'extraction des données.

2.1. *Identification des publications*

Nous avons fait le choix de circonscrire cette revue aux publications de revues scientifiques et de conférences à comité de lecture, sans limitation de date néanmoins. Les publications sont extraites de la base transdisciplinaire Scopus, qui regroupe environ 49000 sources bibliographiques, et présente notamment une grande couverture des Sciences Humaines et Sociales.

Une première étape a consisté à considérer les articles traitant directement du *scenario planning* en recherchant dans leur titre :

– Ses synonymes largement utilisés dans le domaine de la prospective : *scenario method*, *scenario methodology* et *scenario framework*

– Les termes rencontrés dans la littérature pour désigner l'étape d'élaboration des scénarios : *scenario design*, *scenario building*, *scenario construction* et *scenario development*

– Un terme plus générique qui englobe les différentes réflexions devant mener à la sélections des scénarios : *scenario analysis*

Une deuxième étape s'est tournée vers des termes traitant spécifiquement de l'évaluation de la qualité des scénarios (*scenario assessment*, *scenario evaluation*[8] , *scenario*

7. Preferred Reporting Items for Systematic reviews and Meta-Analyses
8. la recherche a été généralisée pour inclure les formes adjectivées (i.e. *scenario assessed* et *scenario evaluated*)

quality) pour s'assurer de constituer un éventail le plus exhaustif possible des articles d'intérêt.

Chacune de ces recherches documentaires est conditionnée par la présence du terme criterion/criteria.

Deux recherches successives sont donc formulées dans la base Scopus :

[Recherche #A] TITLE("scenario building") OR TITLE("scenario design") OR TITLE("scenario development") OR TITLE("scenario construction") OR TITLE("scenario planning") OR TITLE("scenario analysis") OR TITLE("scenario method*") OR TITLE("scenario framework") OR TITLE("prospective scenario") AND TITLE-ABS("criter*")

[Recherche #B] (TITLE-ABS(”scenario assess*”) OR TITLE-ABS(”scenario eval*”) OR TITLE-ABS(”scenario quality”)) AND TITLE-ABS(”criter*”)

Dans la suite du document, les deux groupes d'articles identifiés par les recherches #A et #B seront respectivement nommés groupe #A et groupe #B.

2.2. *Sélection des articles*

Chaque groupe de publications identifiées suit un processus de sélection identique. Premièrement, une étape dite de *consultation* s'attache à supprimer les doublons ainsi que les documents ne répondant pas aux critères d'éligibilité. Dans le cas de cette revue, il s'agit d'écarter les documents qui ne seraient ni des articles de revue, ni des publications de conférences à comité de lecture. En effet, ces *mauvais types* de document (livres, rapports) peuvent être présents dans la sélection malgré les filtres de recherche utilisés dans Scopus. Deuxièmement, une étape de *sélection* se base sur la lecture du titre et du résumé de la publication, et s'appuie principalement sur deux constats.
D'une part, le scénario peut désigner un élément méthodologique très éloigné d'une démarche prospective. En particulier, certains domaines ont fait leur acception propre du terme :

– On parle de scénario en ingénierie en tant que cas d'usage (*use-case* ou *user-story*) utilisé pour éliciter les exigences requises pour la conception de systèmes (Achour, 1998), produits, objets ou services.

– Dans le domaine médical, les scénarios désignent des cas cliniques réalistes utilisés pour les formations de santé (Der Sahakian, 2024). Des scénarios sont également utilisés dans le domaine de la certification pour évaluer le risque lié à la survenue d'un événement critique (incendie (Bjelland and Borg, 2013), séisme (Praticò *et al.*, 2022)). Dans les deux cas, il s'agit d'établir des situations de référence, basées sur les expériences passées, et non d'anticiper des évolutions à venir.

– Une volonté de réduction de l'impact environnemental de l'activité humaine peut également conduire à devoir choisir parmi différentes stratégies de gestion (des déchets, par exemple (Abdallah *et al.*, 2019)). Des simulations, présentées comme des scénarios, peuvent alors être mises en œuvre pour évaluer les solutions techniques proposées en fonction de critères de performance.

Dans tous ces cas, il s'agit de narratifs prédéfinis, qui n'ont aucun caractère hypothétique ou incertain, et ne décrivent pas de situation future. Ce ne sont donc pas des *scénarios prospectifs*. Les articles correspondant à ces *domaines spécifiques* sont ainsi identifiés et écartés de la sélection.

D'autre part, certains articles évoquent des *critères décisionnels* qui ne sont pas utilisés pour évaluer des scénarios. Ces critères permettent en fait d'effectuer des choix stratégiques qui s'offrent à l'utilisateur en fonction d'impacts financiers, managériaux (Gaspars-Wieloch, 2023) ou techniques. Les articles correspondants sont également considérés comme *hors sujet* pour notre revue, et donc écartés de la sélection. Par ailleurs, certains articles font l'objet d'une *sélection erronée* par Scopus (un signe de ponctuation non détecté entre deux termes de recherche, notamment). Enfin, *divers* articles traitent du sujet de façon trop vague, sans détailler de méthode, d'outil ni de cas d'application, et ne s'intéressent à aucun critère spécifique pour l'évaluation de scénarios. Tous ces articles sont également écartés de la sélection.

Les deux sélections obtenues sont ensuite fusionnées pour former un seul corpus.

2.3. *Extraction des données*

Les critères sont identifiés au cours d'une lecture intégrale des articles sélectionnés. On relève les termes utilisés quelle que soit leur nature (nom ou adjectif). A l'issue de la lecture de l'ensemble du corpus d'articles, une vérification est effectuée en lançant une détection automatique de chaque critère listé précédemment dans l'ensemble des textes des articles. A la lecture d'un article, un critère est retenu s'il participe à l'évaluation des scénarios. Une seule occurrence est retenue par article pour un critère donné, même si ce critère est présent plusieurs fois dans l'article. L'occurrence du critère est comptabilisée même si le critère est cité d'une publication antérieure, sauf si cette référence a été publiée par le même auteur principal. On répertorie et on comptabilise également les définitions des critères. Est considéré comme définition tout énoncé permettant de caractériser la nature du critère ou d'en mesurer les propriétés essentielles.

2.4. *Risques de biais transversal aux études*

Des biais ont été identifiés qui pourraient remettre en question le caractère exhaustif de la liste de critères obtenus à l'issue de la revue. sur la méthode d'extraction des critères.

Le premier biais porte sur le protocole de constitution du corpus d'articles.Ce biais concerne le terme *critère (criterion)* utilisé dans la recherche Scopus : deux raisons pourraient en effet expliquer son absence dans un article pourtant pertinent. D'une part, des synonymes du terme pourraient être utilisés. Néanmoins, parmi ceux cités par le Cambridge Dictionary[9], aucun n'a été jugé suffisamment proche d'un point de vue sémantique pour l'intégrer à la recherche dans Scopus.[10]
D'autre part, des articles peuvent s'attacher à l'analyse plus spécifique d'un ou plusieurs critères sans les identifier comme tels.

L'autre biais porte sur la méthode d'extraction des données. Il concerne les définitions proposées dans certains articles, qui sont en fait citées de références antérieures. Même si ces références ne répondent pas aux critères de recherche initiaux, il semble légitime qu'elles puissent apparaître dans le corpus final d'articles, en tant que sources initiales de la définition des critères. Par ailleurs, il existe aujourd'hui une très grande variété de types de scénarios en fonction des cas d'étude envisagés (Crawford, 2019), et mieux saisir le contexte d'emploi d'un critère, dans l'un ou l'autre des articles, peut aider à mieux en comprendre une définition qui y serait proposée.

Pour pallier ces biais, une phase annexe de revue d'articles et d'extraction de critères est menée.

2.5. *Phase annexe de revue*

La phase de revue annexe a consisté à enrichir la sélection initiale par effet cumulatif (*snowballing*), en y intégrant certains des articles qui étaient cités comme sources de définitions des critères. Une étape d'extraction de nouveaux critères et de leurs définitions a ensuite été logiquement entreprise sur ces nouveaux corpus d'articles.

La figure 1 présente les différentes étapes de ce processus de revue annexe, qui a pour point de départ le corpus initial d'articles, ainsi que la liste initiale des critères en ayant été extraits.

1) **Identification** : on identifie les publications citées comme sources de définitions de critères de la liste initiale.

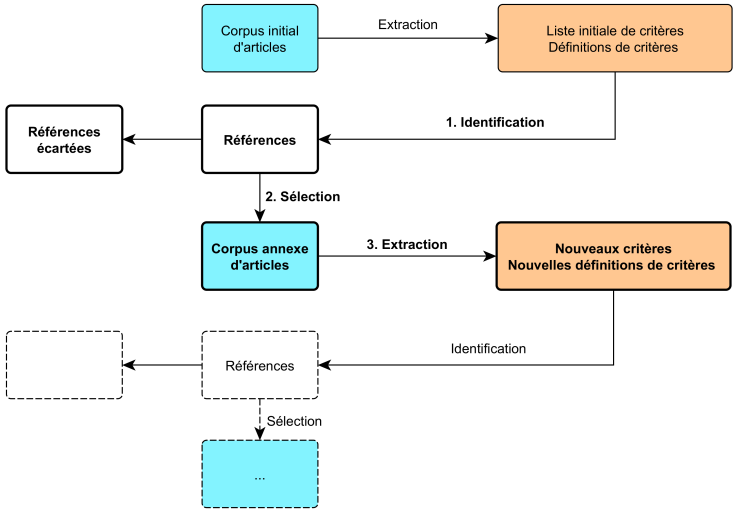
2) **Sélection** : on effectue une sélection de ces publications, selon des conditions similaires à celles utilisées dans la première phase de sélection :

- les publications doivent être des articles de revue ou de conférence

- les publications doivent traiter de l'élaboration ou de la sélection de scénarios dans une démarche de *scenario planning*. Pour cela, on s'assure que l'une des expressions contenant le terme *scenario* utilisées dans les formulations de recherche #A et #B du §2.1 est présente dans le titre ou l'abstract de la publication.

9. https ://dictionary.cambridge.org/fr/dictionnaire/
10. A titre d'exemple, *standard* relève d'un consensus qui n'est pas recherché ici, et *measure* est trop spécifique à une métrique quantitative d'évaluation

FIGURE 1 : Processus de *snowballing* pour la phase annexe de revue



- la publication ne doit pas déjà être présente dans le corpus initial, pour éviter les doublons.

3) **Extraction** : on extrait de ces nouveaux articles les occurrences et définitions des critères existants, ainsi que de nouveaux critères éventuels.

4) Si certaines des nouvelles définitions trouvées à l'étape précédente d'extraction sont elles-mêmes issues d'articles antérieurs, on renouvelle les étapes d'identification, de sélection, puis d'extraction, jusqu'à ce qu'aucun nouvel article ne soit identifié ou sélectionné.

L'ensemble des articles sélectionnés est intégré au corpus initial pour former le corpus global des articles de la revue.

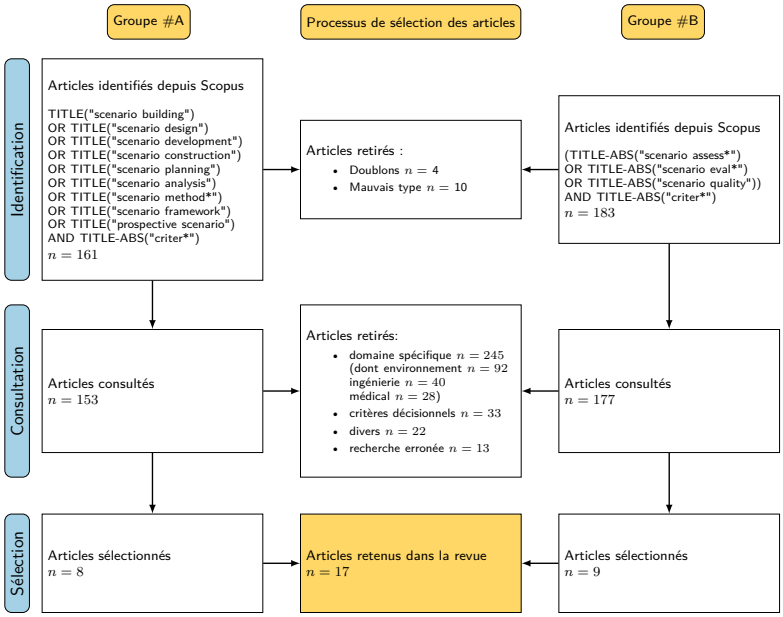
2.6. *Synthèse des résultats*

Deux métriques sont définies pour quantifier les résultats. Ainsi, on comptabilise pour chaque critère :

– le nombre d'occurrences dans la sélection d'articles, c'est à dire le nombre d'articles dans lesquels le critère est cité au moins une fois.

– le nombre de définitions du critère, c'est à dire le nombre d'articles dans lesquels au moins une définition du critère est proposée.

FIGURE 2 : Processus de sélection des articles



La fréquence d'usage d'un critère (mesuré par son nombre d'occurrences) pourrait caractériser une forme de généricité. La présence d'une ou plusieurs définitions indique qu'on a jugé utile, voire nécessaire, de préciser le critère : l'expliciter pour éviter toute ambiguïté, s'accorder sur les moyens de le mesurer. La synthèse s'attache à dégager l'ensemble des critères jugés d'intérêt, selon les valeurs de ces deux métriques.

3. Résultats

3.1. *Sélection des études*

Les résultats chiffrés de la première phase de sélection des articles sont présentés sur la figure 2, qui décrit la sélection des groupes #A et #B avant leur fusion en un corpus de 17 articles.

Au cours de la phase annexe de revue, 65 nouvelles références ont été identifiées comme source de certaines définitions. Parmi ces publications, 43 ont été écartées (25 pour *mauvais type*, 15 pour *hors sujet*, 3 pour *doublon*).

Au total, les deux phases de revue permettent donc de réunir un corpus global de 39 articles. Ces articles sont référencés dans le tableau 1.

TABLEAU 1 : Corpus des articles sélectionnés à l'issue de la revue

Auteurs	**Année**	**Titre**
Jungermann	1985	Inferential processes in the construction of scenarios
Brauers & Weber	1988	A new method of scenario analysis for strategic planning
Bunn & Salo	1993	Forecasting with scenarios
Rotmans et al.	2000	Visions for a sustainable Europe
Heugens & van Oosterhout	2001	To boldly go where no man has gone before: integrating cognitive and physical features in scenario studies
Durbach & Stewart	2003	Integrating scenario planning and goal programming
Van Notten et al.	2003	An updated scenario typology
Swart et al.	2004	The problem of the future: sustainability science and scenario analysis
Bradfield et al.	2005	The origins and evolution of scenario techniques in long range business planning
Rasmussen	2005	The narrative aspect of scenario building - How story telling may give people a memory of the future
Tietje	2005	Identification of a small reliable and efficient set of consistent scenarios
Börjeson et al.	2006	Scenario types and techniques: Towards a user's guide
Liu et al.	2008	Linking science with environmental decision making: Experiences from an integrated modeling approach to supporting sustainable water resources management
Walton	2008	Scanning Beyond the Horizon: Exploring the Ontological and Epistemological Basis for Scenario Planning
Mahmoud et al.	2009	A formal framework for scenario development in support of environmental decision-making
Piirainen & Lindqvist	2009	Enhancing business and technology foresight with electronically mediated scenario process
Wiek et al.	2009	Systemic scenarios of nanotechnology: Sustainable governance of emerging technologies
Wright & Goodwin	2009	Decision making and planning under low levels of predictability: Enhancing the scenario method
Bryant & Lempert	2010	Thinking inside the box: A participatory, computer-assisted approach to scenario discovery
Durance & Godet	2010	Scenario building: Uses and abuses
Nowack et al.	2011	Review of Delphi-based scenario studies: Quality and design considerations
van Vliet et al.	2012	Structure in creativity: An exploratory study to analyse the effects of structuring tools on scenario workshop results
Amer et al.	2013	A review of scenario planning
Comes et al.	2013	An Approach to Multi-Criteria Decision Problems Under Severe Uncertainty
Tourki et al.	2013	Scenario analysis: A review of methods and applications for engineering and environmental systems
Wiek et al.	2013	Plausibility indications in future scenarios

Auteurs	Année	Titre
Kosow	2015	New outlooks in traceability and consistency of integrated scenarios
Lord et al.	2016	Choosing diverse sets of plausible scenarios in multidimensional exploratory futures techniques
Pereverza et al.	2017	Strategic planning for sustainable heating in cities: A morphological method for scenario development and selection
Collier et al.	2018	Scenario Analysis and PERT/CPM Applied to Strategic Investment at an Automated Container Port
Johansen	2018	Scenario modelling with morphological analysis
Schmidt-Scheele	2020	'Plausible' energy scenarios?! How users of scenarios assess uncertain futures
Dhami et al.	2022	Scenario generation and scenario quality using the cone of plausibility
Seeve & Vilkkumaa	2022	Identifying and visualizing a diverse set of plausible scenarios for strategic planning
Yanmaz & Asan	2022	A Fuzzy Measure Based Approach to Scenario Evaluation
Gall et al.	2023	Designing solutions for uncertain futures: a checklist for choosing suitable scenarios
Yanmaz & Asan	2024	A novel scenario planning approach considering criteria interaction in multi-criteria evaluation: An application to urban mobility
Curnin et al.	2025	The Scenario Quality Assessment Method: A New Technique for Verifying the Quality of Scenarios
Yanmaz & Asan	2025	A Novel Approach to Scenario Assessment and Selection

3.2. *Extraction des critères*

94 critères ont été recensés dans le corpus d'articles issus de la première phase de revue. 43 d'entre eux présentaient au moins une définition.
Lors de la deuxième phase de revue, 10 nouveaux critères ont été répertoriés, et 35 critères ont pu être définis qui ne l'étaient pas initialement, grâce aux articles supplémentaires.
Au total, la liste de critères s'élève à 104 termes. Au moins une définition est disponible pour 78 d'entre eux.

3.3. *Synthèse des résultats*

La figure 3 présente les critères répertoriés présentant au moins une définition. La liste est classée selon le nombre de définitions, puis selon le nombre d'occurrences les plus élevés.

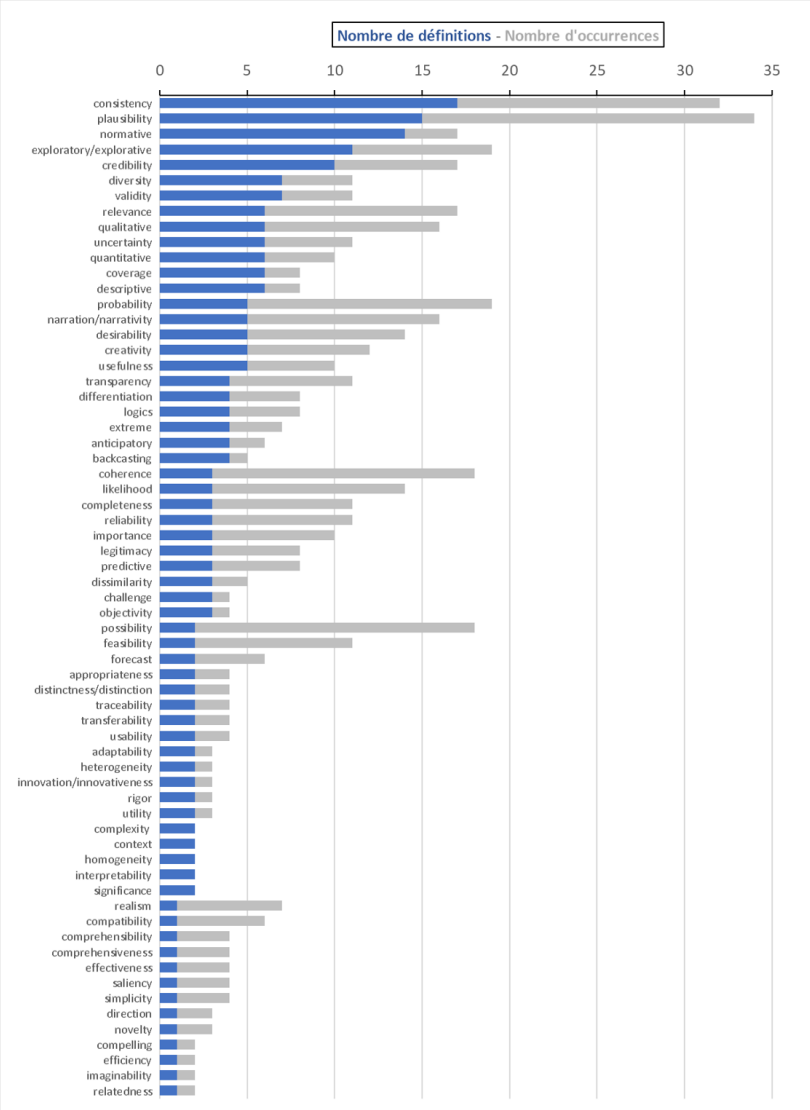


FIGURE 3 : Liste des critères définis

4. Discussion : limites et perspectives

4.1. *Q1 : Liste exhaustive de critères*

La question Q1 avait pour objectif d'établir une liste exhaustive des critères d'évaluation des scénarios.

La démarche suivie dans cette revue a permis d'établir un large panorama des critères considérés et étudiés dans la littérature. Même si leur nombre d'occurrences permet de donner une idée de l'intérêt qui leur est porté, il ne peut pas, toutefois, être considéré comme représentatif de la fréquence de leur usage dans les études prospectives. En effet, un constat demeure des difficultés de mise en pratique de certains de ces critères de façon opérationnelle dans l'évaluation et l'élaboration de scénarios. A titre d'exemple, l'évaluation selon le critère de *consistency*, qui consiste à s'assurer de la cohérence entre les différentes hypothèses considérées dans les scénarios, est souvent considérée comme théoriquement nécessaire (Kosow, 2015), alors qu'elle se heurte en pratique à un problème de combinatoire d'hypothèses trop élevée. (Yanmaz and Asan, 2024).

Par ailleurs, compte tenu de leur très grand nombre, il convient de se demander si tous les critères sont disjoints les uns des autres du point de vue du sens qui leur est donné et de l'utilisation qui peut en être faite. Ainsi, des critères désignés par des termes différents mais représentant la même logique d'évaluation du scénario seraient considérés comme synonymes. Certains articles rapprochent d'ailleurs certains critères en leur donnant une définition commune (*utility/relevance* ou *challenge/novelty* (Amer *et al.*, 2013) par exemple). L'usage de termes différents pour un même signifié pourrait s'expliquer par des applications dans des disciplines différentes, ou par des méthodes de mesure du même critère selon différents types de scénarios à évaluer.
Néanmoins, il n'y a que rarement consensus sur ces rapprochements, comme entre les critères de *consistency* et de *coherence*, souvent considérés comme équivalents (Wiek *et al.*, 2009) (Amer *et al.*, 2013), et pourtant parfois identifiés comme deux notions bien distinctes (Bunn and Salo, 1993).

4.2. *Q2 : Définitions des critères*

La question Q2 s'attachait à identifier les définitions disponibles pour les critères de la liste.

Toutes les définitions ne sont pas comparables en termes de niveau d'explicitation des critères. Elles peuvent consister en un court énoncé (*The degree of being capable of happening* (Yanmaz and Asan, 2025)) ou décrire une métrique complète déjà appliquée pour l'évaluation de cas concrets, comme la Field Anomaly Relaxation (Lord *et al.*, 2016), pour le critère de *plausibility*. Une classification de ces définitions pourrait permettre de distinguer les critères qui font l'objet d'une métrique, et peuvent donc directement s'appliquer à l'évaluation ou l'élaboration d'un scénario, de ceux qui

doivent encore faire l'objet de travaux supplémentaires pour pouvoir être appliqués à des cas concrets.

Par ailleurs, l'intérêt d'un critère n'est pas nécessairement lié à la présence d'articles qui le définissent.
On pourrait souhaiter écarter les critères dont aucune définition n'a été trouvée, et qui paraissent donc difficiles à utiliser pour l'évaluation. Néanmoins, certains comptabilisent de nombreuses occurrences (8 pour *representativeness*, 5 pour *understandability*), ce qui montre que leur usage est malgré tout répandu dans les études référencées. Une recherche plus approfondie concernant ces critères serait nécessaire pour permettre de mieux les caractériser, par exemple en allant chercher dans la littérature au-delà du domaine spécifique du *scenario planning*.
A l'inverse, on peut s'étonner du grand nombre de définitions collectées pour certains critères *courants* (17 pour *consistency*, 15 pour *plausibility*). Le constat a souvent été fait d'un manque de consensus sur la définition des critères (Kosow, 2015), à travers des métriques et des méthodes d'évaluation différentes. Une analyse approfondie de l'ensemble des définitions de ces critères pourrait conduire à proposer des termes différents plutôt que de rester dans l'idée d'un standard qui n'en est pas un.

Enfin, de nombreuses définitions et métriques d'évaluation s'appuient sur celles d'autres critères (*Consistency and plausibility are the decisive conditions for assessing credibility of scenarios.* (Amer *et al.*, 2013)). Ces interdépendances ajoutent à la complexité de l'évaluation des scénarios, en particulier dans la mise en œuvre des méthodes multi-critères évoquées plus haut. Un premier travail de clustering pourrait être envisagé pour identifier ces interdépendances.

4.3. *Conclusion*

Cette revue systématique a été menée dans le but de lister et définir les critères utilisés pour l'évaluation des scénarios au moment de leur élaboration ou de leur sélection. Le processus de revue a suivi les recommandations PRISMA pour rechercher, identifier et sélectionner les articles d'intérêt, puis pour en extraire tous les termes utilisés pour caractériser les scénarios. Deux phases ont été nécessaires pour satisfaire les objectifs : la première a constitué en une recherche ciblée dans la base Scopus, puis une sélection des articles d'intérêt selon un ensemble de critères préétablis ; durant la seconde, la sélection d'articles initiale a été étoffée par enrichissement progressif (*snowballing*) à partir des références citées comme source des définitions des critères. Le corpus global comprend 39 articles de revue et de conférence, d'où 104 critères différents ont été extraits. Pour 78 d'entre eux, au moins une définition a été relevée dans le corpus d'articles, qui décrivait parfois des métriques formalisées et appliquées à des cas d'étude concrets.

Des perspectives ont été proposées pour analyser cette large base de critères, que ce soit pour catégoriser les critères, pour préciser les métriques associées à chaque critère, ou pour étudier les interdépendances à l'intérieur de clusters de critères.

Les travaux futurs devront alors se consacrer à déterminer et caractériser, au sein des scénarios, l'ensemble des éléments clés utilisés pour l'évaluation par ces critères. Ce sont ces éléments qui serviront de base à la définition d'un modèle conceptuel de scénario.

Bibliographie

Abdallah M., Shanableh A., Arab M., Shabib A., Adghim M., El-Sherbiny R., « Waste to energy potential in middle income countries of MENA region based on multi-scenario analysis for Kafr El-Sheikh Governorate, Egypt », *Journal of Environmental Management*, vol. 232, p. 58 - 65, 2019. Publisher : Academic Press Type : Article.

Achour C., « Writing and correcting textual scenarios for system design », *Proceedings Ninth International Workshop on Database and Expert Systems Applications (Cat. No.98EX130)*, p. 166-170, August, 1998.

Amer M., Daim T. U., Jetter A., « A review of scenario planning », *Futures*, vol. 46, p. 23-40, 2013.

Berger G., « L'Attitude Prospective », *Management International*, vol. 4, n° 3, p. 43-46, 1964. Publisher : Springer.

Bjelland H., Borg A., « On the use of scenario analysis in combination with prescriptive fire safety design requirements », *Environmentalist*, vol. 33, n° 1, p. 33 - 42, 2013. Publisher : Kluwer Academic Publishers Type : Article.

Bunn D. W., Salo A. A., « Forecasting with scenarios », *European Journal of Operational Research*, vol. 68, n° 3, p. 291-303, 1993.

Burt G., « Why are we surprised at surprises ? Integrating disruption theory and system analysis with the scenario methodology to help identify disruptions and discontinuities », *Technological Forecasting and Social Change*, vol. 74, n° 6, p. 731-749, July, 2007.

Cordova-Pozo K., Rouwette E. A. J. A., « Types of scenario planning and their effectiveness : A review of reviews », *Futures*, vol. 149, p. 103153, May, 2023.

Crawford M. M., « A comprehensive scenario intervention typology », *Technological Forecasting and Social Change*, vol. 149, p. 119748, December, 2019.

Der Sahakian G. e. a., « The 2024 French guidelines for scenario design in simulation-based education : manikin-based immersive simulation, simulated participant-based immersive simulation and procedural simulation », *Medical Education Online*, 2024. Publisher : Taylor and Francis Ltd. Type : Article.

Gaspars-Wieloch H., « Scenario planning as a new application area for TOPSIS », *Operations Research and Decisions*, vol. 33, n° 2, p. 23 - 34, 2023. Publisher : Wroclaw University of Science and Technology Type : Article.

Girod B., Wiek A., Mieg H., Hulme M., « The evolution of the IPCC's emissions scenarios », *Environmental Science & Policy*, vol. 12, n° 2, p. 103-118, April, 2009.

ICAO/CAEP, Report on the feasibility of a long-term aspirational goal (LTAG) for international civil aviation CO2 emission reductions, Technical report, ICAO International Civil Aviation Organization - CAEP Committee on Aviation Environemntal Protection, March, 2022.

Julien P.-A., Lamonde P., Latouche D., « La méthode des scénarios », , vol. 51, n° 2, p. 253-281, June, 1975.

Kosow H., « New outlooks in traceability and consistency of integrated scenarios », *European Journal of Futures Research*, 2015.

Lord S., Helfgott A., Vervoort J. M., « Choosing diverse sets of plausible scenarios in multidimensional exploratory futures techniques », *Futures*, vol. 77, p. 11-27, 2016.

Page M. J. e. a., « The PRISMA 2020 statement : an updated guideline for reporting systematic reviews », *Croatian Medical Association and School of Medicine*, vol. 57, nº 4, p. 444-465, 2021.

Praticò L., Bovo M., Buratti N., Savoia M., « Large-scale seismic damage scenario assessment of precast buildings after the May 2012 Emilia earthquake », *Bulletin of Earthquake Engineering*, vol. 20, nº 15, p. 8411 - 8444, 2022. Type : Article.

Schoemaker P., « When and how to use scenario planning : A heuristic approach with illustration », , vol. 10, nº 6, p. 549-564, 1991.

Taylor D. C. W., « Alternative World Scenarios for a New Order of Nations », 1993.

Tuominen A., Tapio P., Varho V., Järvi T., Banister D., « Pluralistic backcasting : Integrating multiple visions with policy packages for transport climate policy », *Futures*, vol. 60, p. 41-58, August, 2014.

Van Notten P. W. F., Rotmans J., Van Asselt M. B. A., Rothman D. S., « An updated scenario typology », *Futures*, vol. 35, nº 5, p. 423-443, 2003.

Wiek A., Gasser L., Siegrist M., « Systemic scenarios of nanotechnology : Sustainable governance of emerging technologies », *Futures*, vol. 41, nº 5, p. 284-300, 2009.

Yanmaz O., Asan U., « A novel scenario planning approach considering criteria interaction in multi-criteria evaluation : An application to urban mobility », *Decision Science Letters*, vol. 13, nº 2, p. 461-470, 2024.

Yanmaz O., Asan U., « A Novel Approach to Scenario Assessment and Selection », *International Journal of Information Technology and Decision Making*, 2025.

Une notation iStar plus lisible

Préserver la sémantique tout en améliorant l'expérience utilisateur

Paul Arthur Lemarquis, Anouck Chan, Thomas Polacsek

ONERA
2 Av. Marc Pélegrin
31400 Toulouse
Prenom.Nom@onera.fr

RESUME. *Les langages de modélisation conceptuelle jouent un rôle central dans l'ingénierie des systèmes d'information. Si leurs dimensions sémantiques et syntaxiques ont fait l'objet de nombreux travaux, leur dimension graphique demeure relativement peu étudiée, alors même qu'elle conditionne la compréhension des modèles par les parties prenantes. La notation visuelle peut constituer un obstacle pour les non-spécialistes et limiter l'utilisabilité des modèles, en particulier dans des contextes collaboratifs. Dans cet article, nous proposons une refonte de la notation graphique du langage d'ingénierie des exigences orientée buts iStar, sans en modifier les fondements conceptuels. Notre approche s'appuie sur les principes du design d'information, ainsi que sur des travaux relatifs à la qualité des notations visuelles, notamment la transparence sémantique et l'efficacité cognitive. L'objectif est de préserver l'équivalence sémantique du langage tout en améliorant sa lisibilité et son accessibilité pour des utilisateurs non experts. Notre proposition est illustrée par un exemple simple permettant de comparer la notation existante et la proposition.*

ABSTRACT. *Conceptual modeling languages play a central role in information systems engineering. While their semantic and syntactic dimensions have been extensively studied, their graphical dimension remains comparatively underexplored—despite its decisive impact on stakeholders' comprehension of models. Visual notation can constitute a barrier for non-specialists and limit the usability of models, particularly in collaborative contexts. In this paper, we propose a redesign of the graphical notation of iStar, a goal-oriented requirements engineering language, without altering its underlying conceptual foundations. Our approach draws on information design principles as well as research on visual notation quality, notably semantic transparency and cognitive effectiveness. The objective is to preserve the language's semantic equivalence while enhancing its readability and accessibility for non-expert users. Our proposal is illustrated*

through a simple example that enables a direct comparison between the existing notation and the proposed one.

MOTS-CLÉS : *modélisation graphique ; design d'information ; ingénierie des exigences ; iStar ; design d'interface ; modélisation conceptuelle.*

KEYWORDS: *graphical modeling; information design; requirements engineering; iStar; interface design; conceptual modeling.*

1. Introduction

La modélisation conceptuelle joue un rôle central dans la représentation et l'analyse des systèmes complexes. Ce paradigme de modélisation s'appuie sur des notations graphiques afin de transmettre des informations complexes de manière visuelle. Cependant, les langages de modélisation souffrent généralement d'une limitation : leur notation graphique constitue un obstacle pour les non-spécialistes (Bihanic and Polacsek, 2012). Cette difficulté de compréhension de la notation concerne parfois aussi des modélisateurs expérimentés lorsqu'ils sont confrontés à des modèles de grande taille ou à forte complexité.

Si les travaux de recherche dans le domaine de la modélisation ont largement exploré les dimensions sémantiques et syntaxiques des langages, leur dimension graphique, pourtant essentielle à la compréhension, demeure relativement peu étudiée. Or, la compréhension intuitive et cognitive des modèles par des non-spécialistes dépend directement de cette syntaxe visuelle. Sans remettre en cause les fondements des langages existants, cet article vise à montrer, à travers un exemple, que de simples modifications de la représentation graphique peuvent conduire à des améliorations significatives en termes de compréhension et de lisibilité pour des utilisateurs non experts.

Notons que la problématique de la visualisation des modèles n'est pas nouvelle. Elle a été soulevée il y a près de quatre décennies par Larkin et Simon dans leur article fondateur *Why a Diagram is (Sometimes) Worth Ten Thousand Words* (Larkin and Simon, 1987). Un diagramme et un texte peuvent être informationnellement équivalents, en ce sens qu'ils véhiculent les mêmes faits, cependant, un diagramme rend explicites des relations qui demeurent implicites dans un texte, grâce à sa capacité à organiser spatialement les données. Cette efficacité repose néanmoins sur la capacité des lecteurs à exploiter la structure visuelle proposée. Une notation graphique doit donc permettre aux lecteurs de comprendre aisément et intuitivement le système qu'elle représente. Dans ces conditions, les diagrammes peuvent réduire les efforts cognitifs et favoriser la compréhension d'un système. À l'inverse, en l'absence des connaissances nécessaires à l'interprétation d'un diagramme, un modèle peut se révéler peu informatif, voire source de confusion.

Nous proposons ici d'explorer une refonte, guidée par les principes du *design d'information*, de la notation graphique d'un langage de modélisation. Le design d'information est une discipline qui vise à transformer des données complexes ou abstraites en représentations visuelles adaptées à un public donné. Ses origines remontent au XIX^{e} siècle (Tufte, 2001; Friendly and Wainer, 2021). À titre d'exemple, en 1869, Joseph Minard propose une représentation des flux proportionnelle à leur quantité et à leur direction pour illustrer l'évolution des effectifs de la Grande Armée de Napoléon lors de la campagne de Russie[1]. De manière générale, le design d'information cherche à améliorer la lisibilité, l'accessibilité et, *in fine*, l'efficacité de la communication entre les parties prenantes.

L'objectif de cet article est de définir une nouvelle notation graphique pour le langage d'ingénierie des exigences orientées buts (*Goal-Oriented Requirements Engineering*, GORE) iStar (Yu, 1997; Dalpiaz *et al.*, 2016). Dans la continuité des travaux de Larkin et Simon, nous cherchons à préserver l'équivalence sémantique du langage iStar tout en améliorant sa clarté et son utilisabilité. Nous ne proposons donc aucune modification des concepts ni des mécanismes du langage iStar, nous suggérons uniquement une évolution de sa notation graphique, centrée sur la compréhension des utilisateurs. Notre hypothèse est qu'une telle évolution pourrait améliorer significativement l'accessibilité des modèles iStar pour des non-experts. La contribution principale de cet article est donc de proposer une nouvelle notation graphique pour iStar fondée sur les principes du design d'information.

Le reste de l'article est structuré comme suit. La Section 2 présente la problématique de la qualité visuelle des modèles conceptuels et la dificulté à caractériser cette qualité. La Section 4 décrit notre proposition, à savoir une nouvelle notation graphique pour le langage iStar fondée sur les principes du design d'information. La Section 5 illustre concrètement la notation proposée au moyen d'un exemple simple. La Section 3 situe notre approche au regard des travaux connexes portant sur la qualité des notations et la visualisation des modèles. Enfin, la Section 6 conclut l'article et discute les perspectives de recherche.

2. Qualité visuelle des modèles conceptuels

Malgré l'importance des modèles conceptuels dans l'ingénierie des systèmes d'information, leur qualité visuelle, c'est-à-dire leur lisibilité, demeure souvent négligée. Cette négligence peut engendrer des difficultés de compréhension, d'appropriation et de collaboration entre les parties prenantes. En outre, l'absence de consensus sur ce qui rend un modèle véritablement compréhensible constitue un défi méthodologique majeur.

Comme le soulignent Bork *et al.* (Bork *et al.*, 2018), si les spécifications des langages de modélisation font l'objet d'attention, les techniques de spécification des notations graphiques restent largement sous-investiguées. La définition d'un langage se

1. https://commons.wikimedia.org/wiki/File:Minard.png

concentre généralement sur sa syntaxe abstraite, incluant la formalisation des métaconcepts, de leur sémantique et des relations autorisées entre eux. Bien que ces dimensions soient essentielles, en particulier pour les langages de modélisation visuelle, le choix de la représentation graphique, la syntaxe concrète, est souvent laissé à l'intuition des concepteurs ou aux pratiques établies, plutôt qu'à une analyse systématique et fondée empiriquement.

Plusieurs travaux ont toutefois tenté de structurer l'évaluation de la qualité des notations visuelles. Le cadre proposé par Moody (Moody, 2009), notamment à travers la théorie des *"Physics of Notations"*, identifie neuf principes de conception destinés à améliorer l'efficacité cognitive des notations graphiques. Parmi eux, la transparence sémantique désigne le degré selon lequel la signification d'un symbole peut être déduite de son apparence. Une notation est dite sémantiquement transparente lorsque la forme des symboles suggère intuitivement leur contenu conceptuel. Ce principe présente des affinités avec la notion d'*affordance* en interaction homme-machine, bien qu'il s'en distingue : l'affordance met l'accent sur les actions rendues possibles par un artefact, tandis que la transparence sémantique concerne principalement la facilité d'interprétation des concepts représentés.

Ces considérations rejoignent également la sémiologie graphique de Jacques Bertin (Bertin, 2005). Selon Bertin, un diagramme vise à transmettre visuellement une ou plusieurs informations à l'aide de variables graphiques structurées. L'utilisateur perçoit ces variables selon deux dimensions complémentaires : les variables dites rétiniennes (taille, valeur, texture, couleur, orientation, forme), qui influencent directement la perception visuelle, et les mécanismes liés aux mouvements oculaires, qui participent à l'exploration spatiale du diagramme. Chaque variable graphique possède un pouvoir distinct en termes d'organisation, d'association, de dissociation et de sélectivité, ce qui en fait un levier central pour la conception de notations efficaces.

Ainsi, la qualité visuelle d'un modèle ne relève pas d'une considération esthétique ; elle constitue un facteur déterminant de son efficacité cognitive et communicationnelle. Repenser la notation graphique d'un langage de modélisation suppose donc de mobiliser explicitement ces principes afin d'améliorer la compréhension, en particulier pour des utilisateurs non experts.

3. Travaux connexes

La problématique de la visualisation des modèles conceptuels n'est pas nouvelle (Larkin and Simon, 1987; Moody and Shanks, 1994). Bien qu'il ne s'agisse pas d'un thème de recherche particulièrement en vogue, des travaux sur ce sujet apparaissent de manière relativement régulière. Nous ne cherchons pas ici à être exhaustifs, mais plutôt à présenter quelques contributions, plus ou moins récentes, représentatives du domaine.

En 2015, Gulden et Reijers (Gulden and Reijers, 2015) mettent en évidence une lacune majeure dans la recherche en modélisation conceptuelle : la sous-théorisation de la

conception visuelle. Ils soutiennent que les dimensions perceptives et conceptuelles de la modélisation devraient être étroitement intégrées, en s'appuyant sur des apports issus des sciences cognitives et du design graphique. Alors que les notations de modélisation se concentrent largement sur la sémantique formelle, la syntaxe visuelle est souvent reléguée à un rôle secondaire, voire purement esthétique. Les auteurs proposent un agenda de recherche visant à développer de nouveaux fondements conceptuels unifiant les aspects perceptifs et sémantiques de la modélisation. En dépassant les correspondances traditionnelles univoques entre symboles et concepts, ils appellent à l'élaboration de modèles plus riches de motifs cognitivement efficaces, tenant compte de la manière dont les humains perçoivent, comprennent et interagissent avec les modèles.

McBrien et Poulovassilis (McBrien and Poulovassilis, 2018) proposent une approche de visualisation des données fondée sur des schémas, visant à combler le fossé entre les tables de données de bas niveau et la compréhension conceptuelle des experts du domaine. L'introduction de patrons de schémas de visualisation met en évidence l'intérêt de l'abstraction pour améliorer l'accessibilité et l'interprétabilité d'informations complexes. En établissant un lien entre schémas conceptuels et schémas visuels, cette approche permet de produire des visualisations plus pertinentes et contextualisées, sans exiger des utilisateurs qu'ils manipulent directement des tables de données brutes.

Des travaux plus récents proposent une taxonomie exhaustive de techniques avancées de visualisation de l'information, spécifiquement adaptées aux outils de modélisation conceptuelle (Bork and De Carlo, 2023; Carlo *et al.*, 2022). Cette taxonomie de l'interaction et des données, fournit un cadre orienté conception pour guider le développement d'environnements de modélisation plus intuitifs et interactifs. Elle est construite à partir d'analyses empiriques d'outils de modélisation existants, et recense des mécanismes tels que le zoom, les vues d'ensemble avec détails ou encore les approches focus-contexte. Le résultat de ce travail est un cadre complet pour la classification et pour la création d'outils de modélisation. Si leur approche est large et centrée sur les outils, notre travail se concentre sur la notation d'un langage de modélisation spécifique, à savoir iStar. Notre contribution s'inscrit néanmoins pleinement dans leur appel en faveur de représentations de modèles plus riches et centrées sur l'utilisateur.

Très proche de notre démarche, Moody *et al.* évaluent empiriquement la syntaxe visuelle du langage de modélisation iStar à l'aide du cadre théorique de la *Physics of Notations*, qui regroupe neuf principes visant à concevoir des notations visuelles cognitivement efficaces (Moody *et al.*, 2009). Les auteurs identifient d'importantes faiblesses dans la représentation visuelle d'iStar, notamment : une surcharge symbolique ; une faible discriminabilité perceptive et une complexité graphique élevée. L'ensemble des ses faiblesses nuisent à la compréhension des modèles, en particulier pour les parties prenantes non expertes. À travers une expérimentation contrôlée, l'étude montre que ces problèmes de syntaxe visuelle affectent négativement l'interprétation des modèles et la communication entre acteurs.

Toujours dans le cadre d'iStar, *Gonçalves et al.* proposent une approche empirique pour traiter le problème de *surcharge de symboles* (Gonçalves *et al.*, 2020). Par surcharge de symboles, les auteurs entendent l'utilisation d'un même symbole graphique

pour représenter plusieurs concepts distincts. Les auteurs s'appuient sur une revue systématique de la littérature afin d'identifier des cas de surcharge dans des extensions existantes du langage iStar et se focalisent sur trois extensions. Ils proposent notation basée sur la *Physics of Notations*, ainsi que trois autres notations basées sur des expérimentations empiriques. Leur démarche repose sur une série d'expérimentations avec des étudiants. Les résultats montrent que les symboles conçus à partir de ces études empiriques, ou conçus à partir de la *Physics of Notations*, surpassent ceux proposés par les concepteurs des extensions. Il n'y a cependant pas de meilleurs résultat entre notations réalisée à l'aide de leur procédure empirique et notation réalisée à l'aide de la *Physics of Notations*. Cette contribution met en évidence l'importance d'intégrer des méthodes centrées utilisateur dans la conception de notations graphiques, afin d'améliorer leur clarté et leur adoption.

4. Notre proposition

Afin de faciliter une transition fluide pour les utilisateurs actuels d'iStar, nous avons délibérément choisi de conserver certains éléments graphiques essentiels, tels que les formes représentant les éléments intentionnels, et de rester proches des symboles existants. Par ailleurs, une contrainte pragmatique a influencé l'ensemble de nos décisions de conception : le modèle doit être lisible en noir et blanc. Même lorsque un modèle est imprimé en noir et blanc, sans couleur, notre notation graphique doit rester pleinement fonctionnelle[2]. Tout recours à la couleur est considéré ici comme complémentaire et non essentiel.

Dans cette section, nous présentons en détail notre nouvelle notation graphique pour iStar, ainsi que les motivations et les justifications de nos choix de révision.

4.1. *Éléments intentionnels (les boîtes)*

Les éléments intentionnels représentent ce que les acteurs souhaitent atteindre ou obtenir. iStar distingue les types d'éléments intentionnels suivants (Dalpiaz *et al.*, 2016) :

– **But (*goal*)** : un état du monde que l'acteur cherche à atteindre et dont les critères de satisfaction sont clairement définis.

– **but souple (*soft goal*)** : un attribut pour lequel l'acteur vise un certain niveau de réalisation. Ce niveau peut être précisément spécifié ou rester volontairement vague.

– **Tâche (*task*)** : une action que l'acteur souhaite voir exécuter, généralement pour atteindre un ou plusieurs buts.

2. Nous justifions ce choix par le fait que l'impression en noir et blanc reste courante. De plus, les articles, qu'ils soient publiés dans des actes de conférences ou dans des journaux, peuvent également être soumis à cette contrainte.

– **Ressource** : une entité physique ou informationnelle nécessaire à l'exécution d'une tâche par l'acteur.

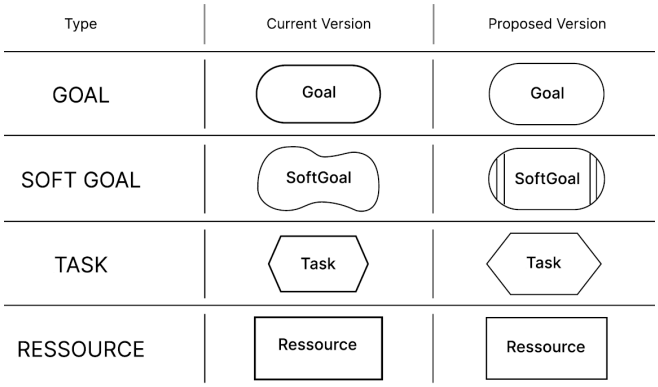


FIGURE 1. *Éléments intentionnels*

Les éléments intentionnels dans la notation iStar présentent une cohérence structurelle, à l'exception notable des buts souples. La forme actuelle en *cacahuète* pour les buts souples est trop distincte de celle des goals, bien qu'ils soient conceptuellement liés (Figure 1). Ceci est un bon exemple d'utilisation du principe d'homographie pour rapprocher symboliquement les deux objets. La proposition est que le but souple reprend la forme du goal, mais avec deux barres encadrant le texte. Cette distinction visuelle renforce la cohérence tout en restant facilement identifiable. Elle facilite aussi la modélisation manuelle, en permettant de créer d'abord tous les goals, puis d'annoter les buts souples.

La conception de notre notation est guidée principalement par l'impératif de préserver la cohérence formelle pour des concepts sémantiquement équivalents. De plus, la prise en compte des contextes de modélisation manuelle a joué un rôle crucial. En particulier, dans les modèles dessinés à la main, les utilisateurs peuvent représenter efficacement tous les buts en utilisant une forme de base homogène et annoter ensuite les buts souples simplement en ajoutant les barres verticales.

4.2. *Relations entre éléments (les flèches)*

Les éléments intentionnels d'iStar sont reliés par différents types de relations, notamment le raffinement, la dépendance, la contribution et la relation *NeededBy*.

Le raffinement est une relation hiérarchique générique reliant des buts ou des tâches. Il s'agit d'une relation n-aire associant un élément parent à un ou plusieurs enfants, un élément ne pouvant être parent que dans une seule relation de raffinement. Deux formes de raffinement existent : le raffinement *ET*, où la satisfaction de tous les enfants

est nécessaire pour satisfaire le parent, et le raffinement *OU inclusif*, où la satisfaction d'au moins un enfant suffit, y compris dans le cas d'un enfant unique.

Les dépendances modélisent des relations sociales entre acteurs. Une dépendance exprime le fait qu'un acteur repose sur un autre acteur pour la fourniture d'un élément, appelé *dependum*.

Les liens de contribution représentent l'influence de buts ou de tâches sur des buts souples. Ils jouent un rôle central dans l'aide à la décision en permettant d'évaluer et de comparer des alternatives. Ces relations traduisent l'accumulation d'indices positifs ou négatifs en faveur de la satisfaction d'un but souple. Les contributions peuvent être positives fortes (*Make*), positives faibles (*Help*) ou négatives faibles (*Hurt*).

Enfin, la relation *NeededBy* relie une tâche à une ressource et indique que celle-ci est nécessaire à l'exécution de la tâche, sans préciser la nature exacte de ce besoin.

Une problématique importante du langage iStar est celle liée à l'homographie des relations (des flèches). Une homographie se produit lorsque le même symbole visuel (mêmes forme, couleur ou lien graphique) est utilisé pour représenter plusieurs concepts ou relations différents, ce qui génère de l'ambiguïté. Dans iStar, un même type de flèche peut représenter différentes relations sémantiques, ce qui nuit à la clarté du diagramme.

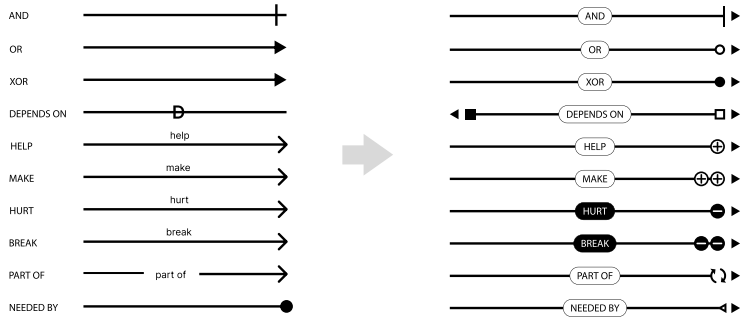


FIGURE 2. *Symbolique proposée*

Pour remédier à ce problème, nous proposons une notation où aucun symbole ne peut être confondu avec un autre porteur d'une signification différente (Figure 2). Dans notre proposition, les relations conservent une cohérence grâce aux flèches, tandis que chaque objet est différencié par un symbole unique. Cela maintient l'unité visuelle tout en assurant une distinction claire des significations.

À cela, nous avons rajouté des balises intégrées directement aux lignes qu'elles dénotent, afin de résoudre le problème fréquent de déconnexion visuelle dans les modèles complexes. Cette intégration réduit l'effort cognitif et améliore encore un peu plus la clarté visuelle. Les balises servent de repères secondaires, décrivant sans définir, ce qui permet à la structure graphique de rester le principal vecteur de sens.

Cette redondance visuelle renforce la compréhension, notamment pour les utilisateurs peu familiers avec le langage.

Nous allons maintenant voir plus en détail certaines relations.

4.2.1. *Contributions positives et négatives*

Les contributions positives et négatives font référence à la manière dont certaines actions, tâches ou buts influencent d'autres buts. La visualisation actuelle utilise uniquement une symbolique textuelle pour différencier les différents types de contribution.

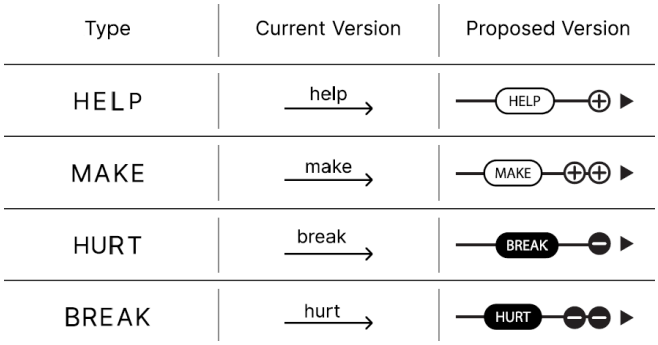


FIGURE 3. *Contributions positives et négatives*

Nous proposons une différenciation plus marquée à travers des symboles uniques pour chaque élément. Nous cherchons ici à exploiter l'homographie, qui dans la même optique d'unification symbolique des éléments en fonction de leur famille sémiotique que pour les éléments intentionnels, permet de solidifier les différentes significations des liens (Figure 3).

La contribution de rupture *(ie break)*, qui indique un impact négatif plus fort, est représentée par un double symbole moins « - - » avec un fond rempli. La contribution négative, qui indique une relation négative plus faible, est représentée par un simple signe moins « - ». Cette différenciation visuelle réduit l'ambiguïté et aligne la force du symbole sur le poids sémantique de la contribution. Comme précédemment, cette approche privilégie la redondance graphique pour améliorer la lisibilité, en particulier pour les utilisateurs novices. La répétition du libellé textuel directement sur la ligne facilite la comparaison entre éléments et réduit la complexité du modèle (Tufte, 2001).

La contribution positive suit les mêmes principes de conception, avec quelques modifications importantes. L'expression « *help* » est remplacée par un signe « + » pour indiquer une influence positive.

La Figure 4 présente un exemple illustratif de la différence entre notre notation et la notation iStar. Ici les buts *Avoir de bons encadrants* et *Le stage a un rapport avec ma passion* ont un impact positif sur le but *Trouver un bon stage* matérialisé

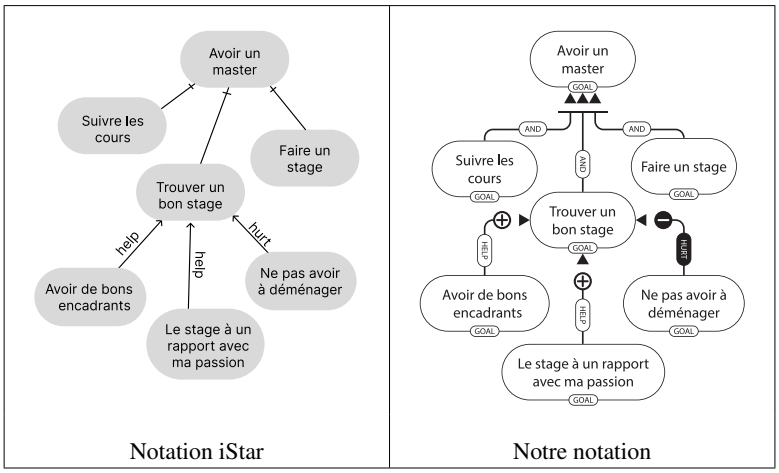


FIGURE 4. *Exemple relation de contribution*

par des flèches *help*. A contrario, le but *Ne pas avoir à déménager* a potentiellement un impact négatif matérialisé par une flèche *hurt*. Cet impact négatif est motivé par le fait que ne pas vouloir déménager est une contrainte qui vient réduire les chances de trouver un bon stage. Nous voyons également sur ces deux diagrammes que la notation graphique iStar ne fait pas de réelle différence notionnelle entre les relations là où notre notation propose une vraie différence, immédiatement perceptible, entre ce qui contribue positivement et ce qui contribue négativement. Toujours Figure 4, nous voyons apparaître une relation de raffinement pour le but *Avoir un master*. Nous allons maintenant voir plus en détail la notation graphique que nous proposons pour les différents raffinements d'iStar.

4.2.2. *Le raffinement*

Suivant les versions d'iStar, il existe deux raffinements : AND et OR, et parfois un troisième est ajouté : le XOR. Le raffinement AND est une relation qui indique qu'un but, un but souple ou une tâche est atteint seulement si toutes ses sous-composantes le sont. Le raffinement OR indique qu'un but, un but souple ou une tâche peut être atteint en accomplissant au moins une de ses sous-composantes. Le raffinement XOR signifie qu'exactement une des sous-composantes doit être réalisée.

La notation du raffinement AND souffre d'un manque de clarté visuo-graphique (Figure 5). Elle utilise une croix qui, en termes d'affordance, évoque visuellement la négation ou l'échec, ce qui entre en contradiction directe avec la signification logique du raffinement AND. L'apparence graphique suggère l'inverse de la définition AND, ce qui nuit à l'interprétation correcte du modèle. Nous avons donc repensé cette représentation graphique pour aligner l'affordance visuelle avec le sens sémantique de la relation :

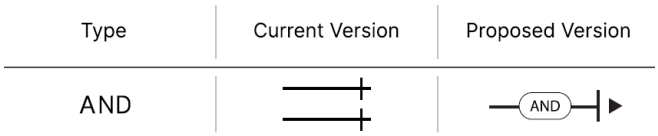


FIGURE 5. *La relation de raffinement AND*

une ligne continue relie les flèches, renforçant l'idée que tous les éléments doivent être satisfaits conjointement (Figure 5). Cette nouvelle représentation, une agrégation des éléments, améliore la lisibilité tout en minimisant le risque de mauvaise interprétation. Une structure plus claire est ainsi suggérée pour simplifier la lecture des flèches.

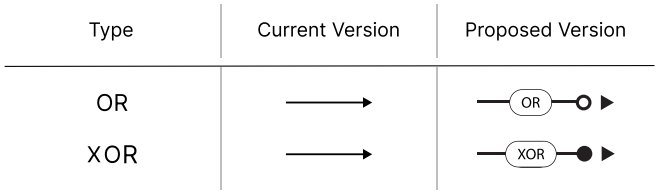


FIGURE 6. *La relation de raffinement Or*

Les raffinements OR et XOR partagent une structure visuelle cohérente, mais reposent sur la séparation. Tous deux utilisent un cercle avant la flèche, en référence au "o" commun dans leur nom (Figure 6). Le cercle est vide pour l'OR, plein pour le XOR, créant un contraste clair qui reflète leur logique respective. Ce rapprochement graphique les unit dans leurs rôles, qui est similaire sans être parfaitement égal. Ils doivent être quand même différenciés, mais de manière discrète.

La Figure 7 présente les raffinements AND et OR suivant la notation iStar, partie gauche, et suivant notre notation, partie droite. Le but *Valider un master* est raffiné en trois sous-buts suivant un raffinement AND. Le but *Prevalider un master* est raffiné un trois sous-buts suivant un raffinement OR. Cet exemple donne à voir, partie gauche de la figure, comment il difficile de faire la différence entre un raffinement AND et un raffinement OR dans la notation iStar.

4.2.3. *La dépendance*

La symbolique utilisée pour définir la dépendance est la lettre "**D**", utilisée de manière directionnelle pour indiquer la direction de la dépendance. Celle-ci prête à confusion, surtout si elle indique une direction contraire à la lecture traditionnelle de la lettre "**D**". Le "**D**" n'est également pas intuitif pour un utilisateur novice. Nous proposons comme solution de réduire cette directionnalité en ajoutant des symboles des deux côtés du lien. La dépendance peut se lire dans les deux sens. La symbolique actuelle est remplacée par des symboles de formes carrées vides ou remplis avec

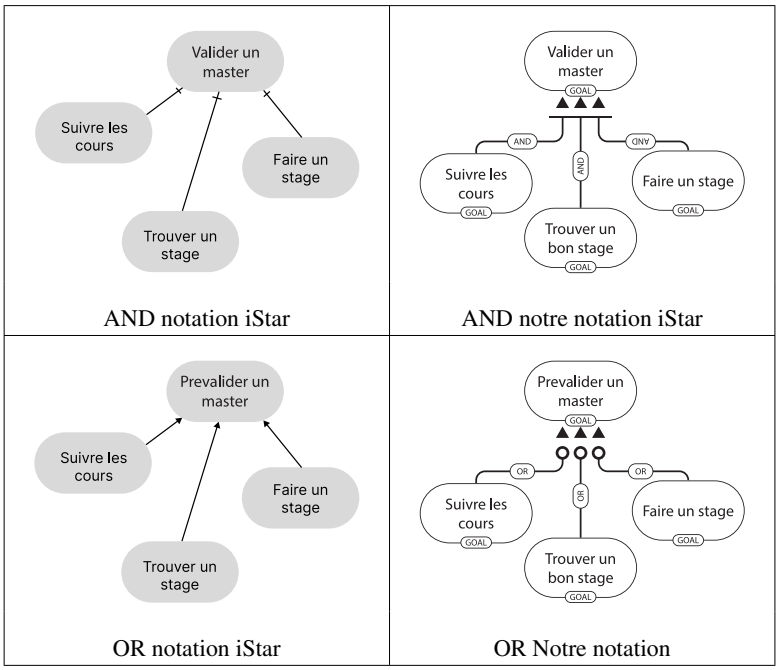


FIGURE 7. *Exemple de raffinement AND et OR*

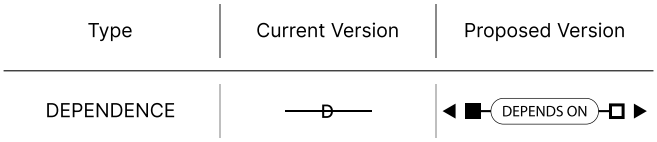


FIGURE 8. *La relation de dépendance*

une pointe allant dans la direction de la dépendance, impliquant la communication entre les différents acteurs (Figure 8). La Figure 9 présente un exemple de relation de dépendance suivant la notation iStar et suivant notre notation.

5. Exemple

Afin de donner à voir plus concrètement les avantages de notre notation, nous avons repris l'exemple de diagramme iStar donné dans le guide du langage : *iStar 2.0*

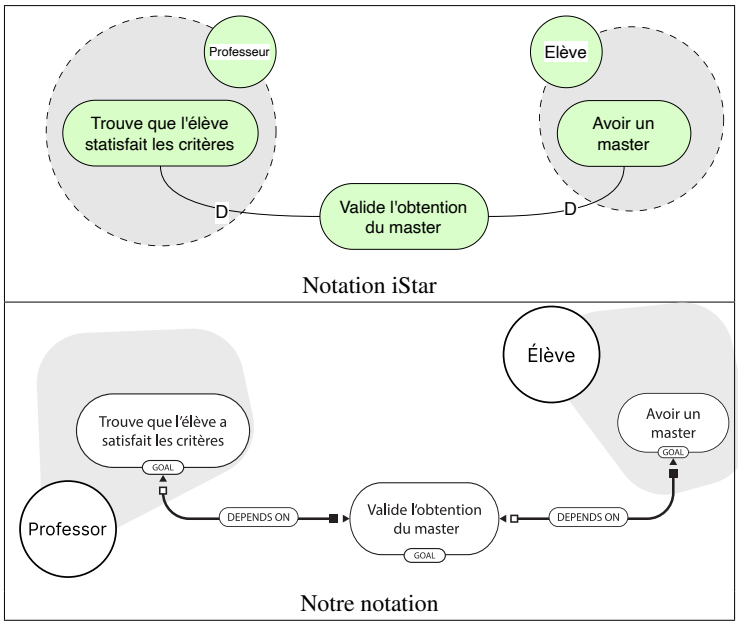


FIGURE 9. *Exemple de dépendance*

Language Guide (Dalpiaz *et al.*, 2016). Plus précisément, nous utilisons le modèle intitulé *A preview of the Travel Reimbursement Scenario* (page 13) pour mettre en évidence les améliorations apportées par notre notation. Les Figure 10 et Figure 11 offrent une comparaison du scénario de remboursement de frais de voyage modélisé selon :

– la notation iStar standard : le diagramme original (Figure 10) ;

– notre notation : le même scénario modélisé à l'aide de notre notation. (Figure 11).

6. Conclusion

Dans cet article, nous avons proposé une évolution de la notation graphique des modèles iStar visant à améliorer leur lisibilité et leur accessibilité, tout en préservant leur sémantique. Notre contribution ne porte pas sur les concepts ni sur la structure du langage lui-même, mais exclusivement sur sa syntaxe concrète, c'est-à-dire sur sa représentation visuelle.

En nous appuyant sur des principes issus du design d'information et sur des travaux relatifs à la qualité des notations visuelles, notamment la transparence sémantique et

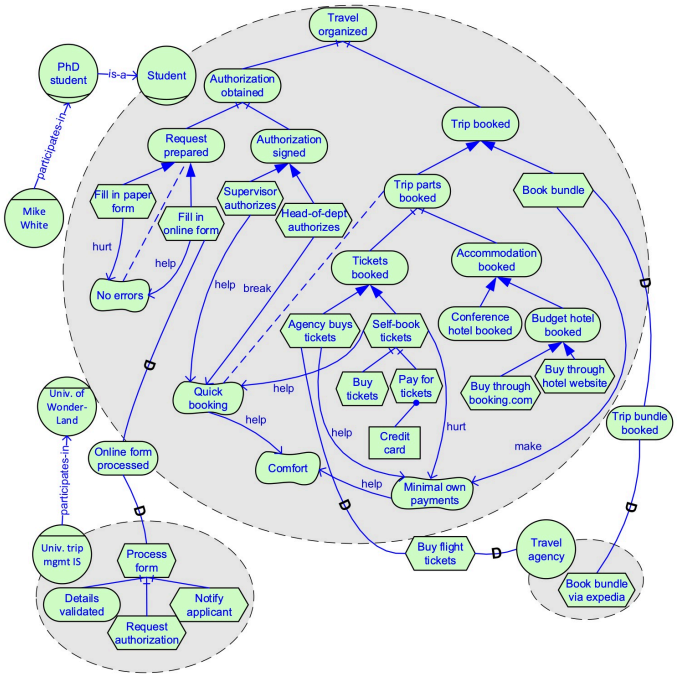


FIGURE 10. *Exemple illustratif utilisant la notation iStar (iStar 2.0 Language Guide, page 13).*

l'efficacité cognitive, nous avons cherché à concevoir une notation plus intuitive et mieux adaptée à des utilisateurs non experts. L'objectif est de réduire les obstacles liés à l'interprétation graphique des modèles et de faciliter la compréhension des relations entre leurs éléments. La proposition a été illustrée au moyen d'un exemple simple, permettant de mettre en évidence les différences structurelles entre la notation existante et celle proposée. Si cette première exploration suggère un potentiel d'amélioration en termes de lisibilité, une validation empirique reste nécessaire afin d'évaluer rigoureusement l'impact de cette nouvelle notation sur la compréhension, la charge cognitive et la collaboration entre parties prenantes. Plus largement, ce travail souligne l'importance de considérer la dimension graphique des langages de modélisation comme un objet de recherche à part entière.

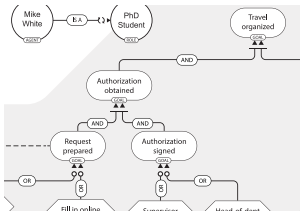


FIGURE 11. *Exemple illustratif utilisant notre notation.*

Bibliographie

Bertin J., *Sémiologie graphique : les diagrammes, les réseaux, les cartes*, Éditions EHESS, Paris, 2005.

Bihanic D., Polacsek T., « Models for Visualisation of Complex Information Systems », *16th International Conference on Information Visualisation, IV 2012, Montpellier, France, July 11-13, 2012*, IEEE Computer Society, p. 130-135, 2012.

Bork D., De Carlo G., « An extended taxonomy of advanced information visualization and interaction in conceptual modeling », *Data Knowl. Eng.*, September, 2023.

Bork D., Karagiannis D., Pittl B., « Systematic analysis and evaluation of visual conceptual modeling language notations », *12th International Conference on Research Challenges in Information Science, RCIS 2018*, IEEE, p. 1-11, 2018.

Carlo G. D., Langer P., Bork D., « Rethinking Model Representation - A Taxonomy of Advanced Information Visualization in Conceptual Modeling », *in* J. Ralyté, S. Chakravarthy, M. K. Mohania, M. A. Jeusfeld, K. Karlapalem (eds), *Conceptual Modeling - 41st International Conference, ER 2022 Proceedings*, vol. 13607 of *Lecture Notes in Computer Science*, Springer, p. 35-51, 2022.

Dalpiaz F., Franch X., Horkoff J., « iStar 2.0 Language Guide », *CoRR*, 2016.

Friendly M., Wainer H., *A history of data visualization and graphic communication*, vol. 56, Harvard University Press Cambridge, 2021.

Gonçalves E. J. T., Almendra C. C., Goulão M., Araújo J., Castro J., « Using empirical studies to mitigate symbol overload in iStar extensions », *Softw. Syst. Model.*, vol. 19, n° 3, p. 763-784, 2020.

Gulden J., Reijers H. A., « Toward Advanced Visualization Techniques for Conceptual Modeling », *in* J. Grabis, K. Sandkuhl (eds), *Proceedings of the CAiSE 2015 Forum at the 27th International Conference on Advanced Information Systems Engineering*, vol. 1367 of *CEUR Workshop Proceedings*, CEUR-WS.org, p. 33-40, 2015.

Larkin J. H., Simon H. A., « Why a Diagram is (Sometimes) Worth Ten Thousand Words », *Cogn. Sci.*, vol. 11, n° 1, p. 65-100, 1987.

McBrien P., Poulovassilis A., « Towards Data Visualisation Based on Conceptual Modelling », *in* J. Trujillo, K. C. Davis, X. Du, Z. Li, T. W. Ling, G. Li, M. Lee (eds), *Conceptual Modeling - 37th International Conference, ER 2018 Proceedings*, vol. 11157 of *Lecture Notes in Computer Science*, Springer, p. 91-99, 2018.

Moody D. L., « The "Physics"' of Notations : Toward a Scientific Basis for Constructing Visual Notations in Software Engineering », *IEEE Trans. Software Eng.*, vol. 35, n° 6, p. 756-779, 2009.

Moody D. L., Heymans P., Matulevicius R., « Improving the Effectiveness of Visual Representations in Requirements Engineering : An Evaluation of i* Visual Syntax », *RE 2009, 17th IEEE International Requirements Engineering Conference*, IEEE Computer Society, p. 171-180, 2009.

Moody D. L., Shanks G. G., « What Makes a Good Data Model ? Evaluating the Quality of Entity Relationship Models », *in* P. Loucopoulos (ed.), *Entity-Relationship Approach - ER'94, Business Modelling and Re-Engineering, 13th International Conference on the Entity-Relationship Approach, 1994, Proceedings*, vol. 881 of *Lecture Notes in Computer Science*, Springer, p. 94-111, 1994.

Tufte E. R., *The Visual Display of Quantitative Information*, second edn, Graphics Press, Cheshire, Connecticut, 2001.

Yu E. S. K., « Towards modelling and reasoning support for early-phase requirements engineering », *Proceedings of ISRE 1997*, IEEE, p. 226-235, 1997.

ETHCOMOD : une nouvelle méthode de modélisation d'ontologie métier

Application dans le secteur de l'énergie

Charlotte Darricades[1,2,3], Christian Sallaberry[2], Sébastien Laborie[2], Eric Kergosien[1], Patrice De La Broise[1]

1. Universite de Lille, Gériico
charlotte.darricades@univ-lille.fr, eric.kergosien@univ-lille.fr, patrice.de-la-broise@univ-lille.fr
2. Universite de Pau et des Pays de l'Adour, E2S UPPA, LIUPPA
christian.sallaberry@univ-pau.fr, sebastien.laborie@univ-pau.fr
3. Pôle d'Études et de Recherche de Lacq, TotalEnergies SE, BP 47, 64170 Lacq
charlotte.darricades@totalenergies.com

RESUME. La méthode ETHCOMOD étend les approches existantes en intégrant étroitement chercheurs, communicants et ressources sémantiques externes. Elle repose sur une démarche itérative combinant enquêtes de terrain, analyses de corpus scientifiques et traitements semi-automatisés du langage. Elle a été expérimentée en entreprise, dans un contexte de R&D multi-énergies. Les résultats montrent que l'ontologie produite facilite la désambiguïsation terminologique et la construction d'un langage commun entre experts et non-experts. Le prototype CommTools, développé comme preuve de concept, met en évidence le rôle médiateur de l'ontologie dans la génération de contenus vulgarisés fiables.

ABSTRACT. The ETHCOMOD method extends existing ontology engineering approaches by tightly integrating researchers, communication professionals, and external semantic resources. It relies on an iterative process combining field studies, scientific corpus analysis, and semi-automated natural language processing. The method was validated in an industrial multi-energy R&D context. The experiment shows that the resulting ontology supports terminological disambiguation and enables the construction of a shared language between experts and non-experts. The CommTools proof-of-concept highlights the mediating role of ontology-based knowledge structures in generating reliable popularized scientific content.

Mots-clés : Modélisation d'ontologie, Représentation de connaissances, Partage de connaissances, Communication

KEYWORDS: Ontology engineering, Knowledge representation, Knowledge sharing, Communication

INFORSID 2026 - Catégorie Recherche

1. Introduction

Dans un environnement marqué par l'accélération des transitions énergétiques, numériques et sociétales, le secteur de la Recherche et Développement (R&D) ne se limite plus à inventer et développer des solutions techniques (Lanciano-Morandat et al., 2019 ; Darbellay, 2012) : il contribue activement à la construction des récits organisationnels, à l'identité des entreprises et à leur inscription dans un horizon collectif de transformation. Autrement dit, la transformation de la R&D en opérateur de récits organisationnels (Giroux & Marroquin, 2005) et de légitimation sociale ne peut se réaliser qu'à travers des dispositifs de communication interne et externe capables d'articuler la production de savoirs avec leur mise en circulation. La question de la communication devient alors centrale : comment ces savoirs et ces récits produits par la R&D circulent-ils dans l'organisation et au-delà de ses frontières ?

Dans ce prolongement, les travaux d'El Mezouari, Lotfi et Bouthir (2013) soulignent que la compétitivité des entreprises industrielles dans un contexte mondialisé ne peut être dissociée d'un système de communication interne et externe efficace. Les chercheurs, traditionnellement perçus comme des producteurs de connaissances techniques, sont aujourd'hui de plus en plus sollicités pour devenir des acteurs de communication. Ils ne se limitent plus à publier des résultats scientifiques dans des revues spécialisées : ils participent à des événements internes, coconstruisent des supports de vulgarisation, interviennent dans les médias et contribuent ainsi à l'humanisation et à la mise en visibilité de l'innovation. Face à eux, les communicants occupent une place spécifique mais complémentaire. Leur rôle ne se réduit pas à « traduire » les savoirs en langage accessible : ils cadrent les messages, organisent les dispositifs éditoriaux et participent à la scénarisation stratégique des projets. Ils opèrent une médiation constante entre les exigences de lisibilité et de visibilité d'une part, et les impératifs de précision et de crédibilité scientifique d'autre part. La R&D apparaît donc non seulement comme un espace de production scientifique, mais aussi comme un nœud communicationnel structurant les interactions entre chercheurs, communicants et directions.

Ainsi, la transmission et la vulgarisation de connaissances scientifiques constituent des enjeux stratégiques majeurs. Pourtant, au sein de grandes organisations en pleine mutation face au défi de la transition énergétique, comme, par exemple, TotalEnergies, une rupture sémantique persiste souvent entre les chercheurs, qui produisent un savoir hautement spécialisé, et les communicants, chargés de valoriser ces travaux auprès de publics non experts. Si de nombreuses méthodologies de structuration de la connaissance existent, telles que Methontology (Fernandez-Lopez et al., 1997), NeOn (Suàrez-Figueroa et al., 2012) ou SAMOD (Peroni, 2016), elles se concentrent principalement sur des aspects informatiques ou de gestion, délaissant souvent la dimension médiatrice de la communication. L'originalité de notre approche se distingue par une modélisation qui consiste plutôt à engager tous les acteurs qui participent à la gestion des connaissances d'une organisation, et cela tout au long du cycle de vie, c'est-à-dire de la création de ces connaissances jusqu'à leur valorisation.

Cet article présente ETHCOMOD, une nouvelle méthode de modélisation d'ontologie métier conçue pour mieux intégrer les experts dans la boucle de l'identification, de la structuration et du partage des connaissances au sein d'une organisation. En s'appuyant sur une approche itérative et collaborative, ETHCOMOD vise à transformer les productions scientifiques en concepts sémantiques structurés pouvant être, par la suite, exploités à des fins de médiation, entre chercheurs et communicants, par exemple. Ainsi, nous détaillerons d'abord, en section 3, les cinq étapes de cette méthode, avant d'exposer un exemple de mise en œuvre concrète au sein du Pôle d'Études et de Recherche de Lacq (PERL) de TotalEnergies, en section 4. Enfin, dans la section 5, nous présenterons et expérimenterons le prototype CommTools qui exploite cette base de connaissance au service des communicants de TotalEnergies, pour un meilleur accès et une valorisation des productions scientifiques des chercheurs.

2. État de l'art

Un nombre important de travaux propose une méthodologie pour construire une ontologie. Parmi ceux-ci, nous pouvons notamment citer les méthodes TOronto Virtual Enterprise (TOVE) (Grüninger et Fox, 1995), Methontology (Fernandez-Lopez et al., 1997), Sensus (Swartout et al., 1997), OnTo-Knowledge (OTK) (Staab and al., 2001), Terminae (Aussenac-Gilles et al., 2008), Networked Ontologies (NeOn) (Suàrez-Figueroa et al., 2012), ou encore Simplified Agile Methodology for Ontology Development (SAMOD) (Peroni, 2016).

Toutes ces méthodes commencent par une phase d'acquisition de connaissances du domaine ou du métier, de rédaction de spécifications fonctionnelles ou de questions de compétence. Methontology et OTK sont très similaires. Les deux approches commencent par l'acquisition de connaissances et la rédaction de spécifications. Elles se poursuivent en modélisant le domaine d'abord d'une manière informelle puis dans un langage formel. Enfin, toutes deux proposent une évaluation de l'ontologie produite. NeOn et SAMOD édictent des consignes pour créer des ontologies modulaires suivant une approche par composition (Despres, 2014). Certaines méthodologies, notamment OTK, NeOn et SAMOD proposent un développement modulaire de l'ontologie construite pas-à-pas, soit en ajoutant, à chaque itération, la modélisation d'une partie supplémentaire du métier/domaine, soit en modélisant toutes les parties du métier/domaine d'abord, puis en les fusionnant. NeOn se distingue des autres méthodologies présentées car elle fournit de nombreuses approches pour élaborer une ontologie ou un réseau ontologique. Elle demande aux ontologues de réaliser préalablement une analyse approfondie du projet afin de pouvoir choisir la bonne combinaison des processus et activités proposés. Les trois méthodes OTK, NeOn et SAMOD intègrent également une phase d'évaluation de l'ontologie produite, et cela lors de l'étape finale. SAMOD semble ressortir du lot selon nos critères car elle propose une première étape permettant de créer un premier squelette d'ontologie à partir des besoins exprimés par les utilisateurs cibles. La méthode préconise d'impliquer fortement les experts concernés afin de préciser et d'étendre les besoins, et le modèle ontologie produit,

de façon itérative (Rawsthorne et al., 2022). L'aspect itératif impliquant les experts est primordial dans un secteur spécifique tel que celui de la R&D. Enfin, SAMOD intègre des phases de tests à différentes étapes du processus.

Le tableau suivant offre une synthèse de ces approches en mettant en avant leurs forces ainsi que leurs limites.

Tableau 1. Comparaison des principales approches de création d'ontologies

Approches	**Objectifs**	**Forces**	**Limites**
TOVE (1995)	Formaliser les connaissances organisationnelles dans les entreprises industrielles	Une approche rigoureuse et très structurée avec un ancrage fort dans la modélisation des processus d'entreprise	Complexe à exploiter car nécessite une expertise en logique formelle (peu adaptée à la collaboration non technique)
Methontology (1997)	Construire une ontologie complète à partir de la connaissance experte d'un domaine	Une méthode détaillée et systématique avec un guide d'évaluation clair	Interactions faibles entre experts et concepteurs
Sensus (1997)	Étendre et relier des ontologies existantes à partir de ressources linguistiques et sémantiques	Favorise la réutilisation de connaissances et l'interopérabilité entre ontologies	Aucune démarche participative et dépendance forte à des connaissances préexistantes
OnTo-Knowledge (2001)	Développer des ontologies pour améliorer la gestion des connaissances dans les entreprises	Une approche orientée entreprise qui prend en compte la maintenance et l'évaluation continue	Nécessite des ressources informatiques et humaines conséquentes ; ce qui accroît sa complexité
Terminae (2008)	Construire une ontologie à partir de corpus textuels et d'analyses linguistiques	Un fort ancrage sémantique qui est utile pour modéliser un vocabulaire métier	Nécessite des outils linguistiques et un corpus bien délimité (peu d'interactions humaines)
NeOn (2012)	Favoriser la création collaborative et modulaire d'ontologies ou de réseaux d'ontologies	Une approche modulaire et flexible qui permet la co-construction entre plusieurs acteurs	Nécessite un haut niveau d'expertise et une planification lourde
SAMOD (2016)	Concevoir une ontologie de manière agile et itérative en collaboration avec les experts du domaine	Une méthode itérative, participative, flexible et centrée sur les besoins réels des utilisateurs	Demande une implication constante des experts du domaine

L'analyse comparative des méthodologies existantes souligne ainsi la diversité des approches possibles, chacune répondant à des finalités spécifiques techniques, organisationnelles ou linguistiques, mais rarement à des enjeux communicationnels. Si ces méthodes offrent des cadres rigoureux pour la formalisation et la structuration des connaissances, elles demeurent majoritairement ancrées dans une logique informatique ou de gestion de projet, laissant en marge les dimensions symboliques, relationnelles et médiationnelles propres aux Sciences de l'Information et de la Communication. C'est pourquoi dans la section suivante, nous présentons notre approche ETHCOMOD pour répondre à ce besoin.

3. La méthode ETHCOMOD

ETHCOMOD (*cf.* Figure 1) étend la méthode SAMOD (Peroni, 2016), illustrée plus haut, de la façon suivante : (i) ajout d'une première étape qui vise la cartographie de l'organisation en général et du métier ciblé en particulier ; (ii) intégration de différentes communautés dans la collecte et la validation des données ; (iii) préconisation de l'usage d'outils de TALN pour l'extraction d'information (labels et concepts) ; (iv) ajout d'une étape d'enrichissement des connaissances à partir de ressources externes ; (v) intégration des *modlets* (i.e., sous graphes ontologiques, agrégés en fin de processus) pour constituer l'ontologie finale qui sera expérimentée par les communautés du domaine dans leurs activités. Un modlet (Peroni, 2016) permet de modéliser une partie (domaine/sous-domaine) du métier visé sous la forme d'un sous-graphe ontologique.

ETHCOMOD (représentation simplifiée, *cf.* Figure 1) se décline en 5 étapes que nous allons présenter succinctement dans les sections 3.1 à 3.5. Par la suite, un exemple dans le secteur de l'énergie illustrera chaque étape en section 4.

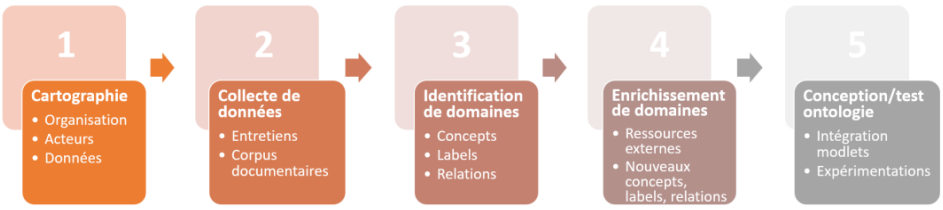


Figure 1 : Les 5 étapes de la méthode ETHCOMOD

3.1. Cartographie de l'organisation

Cette étape consiste à créer une cartographie qui fait état des caractéristiques de l'organisation : directions, acteurs, groupes socio-professionnels, connaissances, informations, données. Plusieurs communautés peuvent émerger (e.g., ingénieurs et chercheurs qui publient des articles scientifiques relatifs à leurs travaux, communicants qui publient également selon différents canaux en interne et en externe pour valoriser des choix/orientations de l'organisation). Les ressources de données créées et/ou gérées par ces communautés sont également répertoriées.

3.2. Collecte de données du domaine (entretiens et corpus textuels)

Il s'agit ici de récolter des données relatives au domaine ciblé. Les approches ethnographiques préconisées (enquêtes, entretiens semi-directifs et observations), menées auprès des communautés en interaction au cœur du métier à modéliser (les vocabulaires peuvent être très différents allant de celui de l'expert jusqu'à celui du néophyte), visent l'identification de corpus documentaires (confidentiels *vs* non-confidentiels) traitant du métier.

3.3. Identification de domaines

Cette étape est consacrée à l'analyse des transcriptions d'interviews et des documents des corpus pour déterminer des sous-domaines et les concepts ainsi que les relations qui les caractérisent. Le regroupement de concepts et des labels associés permet ensuite l'organisation de ces concepts via des liens hiérarchiques. Le traitement de documents semi-automatisé est supporté par des outils de traitement automatique du langage. La validation/correction itérative des sous-domaines représentés par des *modlets* implique des représentants des différentes communautés d'acteurs concernées.

3.4. Enrichissement de domaines

ETHCOMOD fait une hypothèse forte que des ressources externes expertes (bases de connaissances en lien avec un sous-domaine) peuvent enrichir efficacement des *modlets* en termes de concepts, labels et relations. Ici, comme précédemment, des traitements semi-automatisés d'interrogation et de parcours de bases de connaissances sont préconisés.

3.5. Conception et tests de l'ontologie (intégration des modlets, expérimentations)

Les *modlets* sont intégrés et l'ontologie métier est expérimentée/utilisée par les communautés du domaine dans leurs activités. La validation/tests auprès d'experts est réalisée à l'aide d'un éditeur d'ontologie. Des applications informatiques exploitant l'ontologie sont également développées et mises à disposition des utilisateurs.

4. Application de la méthode ETHCOMOD aux métiers de l'énergie

Nous avons choisi le contexte des métiers de l'énergie et plus particulièrement la collecte, la structuration et la vulgarisation des connaissances au sein du PERL (Pôle d'Etudes et de Recherche de Lacq), une entité du groupe TotalEnergies. Il s'agit, notamment, de permettre à deux communautés distinctes de l'organisation de partager des connaissances et des données relatives au domaine de l'énergie, dans leurs activités au service de l'organisation. Ainsi, une communauté d'ingénieurs et de chercheurs mène des activités de R&D dont elle fait la promotion par des

publications scientifiques auprès de pairs dans des conférences et revues nationales et internationales. Elle peut avoir des difficultés à vulgariser ces activités en interne auprès de non spécialistes lors de présentations ou autres comptes-rendus. De même, une communauté de communicants mène des activités de promotion des activités du groupe en interne et en externe. Elle s'appuie notamment sur les résultats scientifiques obtenus par ces ingénieurs et chercheurs. Elle rencontre des difficultés à interroger des documents scientifiques dont le vocabulaire spécialisé peut être difficile d'accès et éloigné du vocabulaire du public ciblé.

Cette expérimentation de la méthode ETHCOMOD a un double objectif : traduire les discours scientifiques en concepts sémantiques structurés, afin d'en permettre la réutilisation et la valorisation par les communicants. Ainsi, il s'agit de créer un langage commun aux chercheurs et communicants, permettant de relier les terminologies expertes à des catégories plus génériques, compréhensibles par le grand public. Notons que, en interne, TotalEnergies dispose d'un thesaurus géologique complet, utilisé pour les métiers du sous-sol, mais aucun référentiel spécifique aux énergies renouvelables ou à la R&D multi-énergies n'existe. De même, en externe, plusieurs ontologies ou thesaurus (VoInrae, UNBIS, PubChem), détaillés dans (Darricades, 2026) couvrent certains domaines (chimie, biologie, environnement, respectivement), mais aucune ne relie ces champs dans une structure interdisciplinaire.

C'est dans ce contexte que nous avons mis en application chaque étape de la méthode ETHCOMOD en collaboration avec les chercheurs et communicants du PERL, chez TotalEnergies.

4.1. Cartographie de l'organisation

Une étude itérative, reliant observations, entretiens, analyses documentaires et interactions de terrain a permis de tracer les bases de la cartographie présentée sur une structure visuelle (voir Darricades, 2026). Les premières investigations ont permis de dégager les lignes hiérarchiques et les pôles décisionnels directement impliqués dans la R&D. En parallèle, une analyse des groupes socio-professionnels a permis de distinguer les chercheurs, les chercheurs-communicants, les communicants et les communicants-technophiles. Les circuits de communication formels et informels, les collaborations inter-équipes et les points de passage entre chercheurs et communicants ont pu être affinés. Un second travail a consisté à recenser, trier et classer l'ensemble des documents produits au sein du centre, en distinguant deux grands ensembles : d'une part les documents confidentiels (cahiers de laboratoire, notes de synthèse, brevets et demandes de brevets, rapport annuels, notes mensuelles de faits marquants de la R&D, les rapports de projets de R&D) et d'autre part, les documents non confidentiels, à savoir les publications scientifiques et communications validées pour diffusion dans un outil interne nommé ROPA (Relation avec les Organismes Professionnels Amont) dédié à la capitalisation des productions scientifiques.

Cette cartographie constitue le premier livrable de la méthode ETHCOMOD, matérialisant les connexions entre directions, individus et groupes professionnels et pointant les ressources informationnelles de chacun.

4.2. Collecte de données du domaine (entretiens et corpus textuels)

La deuxième étape est consacrée à la collecte et à la structuration des données de la R&D et des communicants. Pour notre expérimentation, cette phase s'est concentrée sur la constitution du corpus des publications scientifiques, seules sources publiques, validées institutionnellement et diffusables. Au total, 56 publications produites par les chercheurs du PERL ont été sélectionnées dans le cadre du travail de recensement (sélection représentant les différents domaines de recherche du PERL) via l'outil interne ROPA, avant d'être analysées, classées par thématique et sélectionnées pour la modélisation sémantique. En parallèle, treize entretiens semi-directifs (via grille d'entretien d'une durée moyenne de 1h30) ont été réalisés auprès de sept chercheurs et six communicants du PERL, complétés par des réunions internes et un lien direct avec le terrain à travers quatre stages d'observation. Ces immersions relèvent d'une observation participante visant à saisir les logiques communicationnelles à l'œuvre dans la production et la valorisation scientifique (Latour et Biezunski, 1989). Enfin, un troisième travail intègre une dimension comparative et participative, à travers la diffusion d'un questionnaire (voir Darricades, 2026) : un premier lot, adressé aux communicants R&D internes à TotalEnergies (12 répondants, 9 basés en France, et 3 en Belgique, en Inde et au Brésil) et un second, destiné à des communicants R&D externes à TotalEnergies dans des domaines d'activités stratégiques différenciés (5 répondants français issues des domaines de la construction, pharmaceutique, agroalimentaire et fabrication pneumatiques).

4.3. Identification de modlets (processus par sous-domaine, itérations)

Il s'agit désormais de soumettre les données scientifiques collectées au processus d'extraction terminologique et conceptuel d'ETHCOMOD. Les retranscriptions d'entretiens ont été analysées manuellement par des experts communicants R&D de TotalEnergies. Les 56 publications répertoriées ont, quant à elles, été analysées via une chaine de traitement semi-automatisée. Deux choix fondamentaux ont été opérés : l'utilisation de TextRazor (Hollink et al., 2016) comme outil d'extraction afin d'identifier automatiquement les entités nommées et l'adoption de Wikidata (Vrandečić, 2012) comme référentiel sémantique garantissant la cohérence, la traçabilité et la hiérarchisation des concepts scientifiques (via les identifiants Wikidata QID). Ces choix, parmi de nombreux outils d'extraction d'information et ressources externes, ont été faits à l'issue d'expérimentations sur le corpus (Darricades, 2026).

Ainsi, dans le cadre de cette étude, l'extraction d'information et l'identification de *modlets* à l'étape 3 d'ETHCOMOD se décline en 4 phases :

1) Extraction de termes à forte densité sémantique dans une publication avec TextRazor. L'analyse porte sur les *titre, résumé, introduction et mots-clés*

d'une publication. Ainsi, TextRazor détermine le contexte sémantique approprié dans Wikidata. Ici, le terme *adsorption* est identifié dans une des parties ciblées du texte relatif à une publication du corpus (*cf.* Figure 2) ;

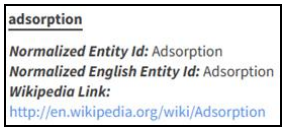


Figure 2 : Extraction de termes

2) Étiquetage conceptuel de termes avec Wikidata. Le terme *adsorption* est rattaché à un concept Wikidata dont le QID est Q180254 (*cf.* Figure 3) ;

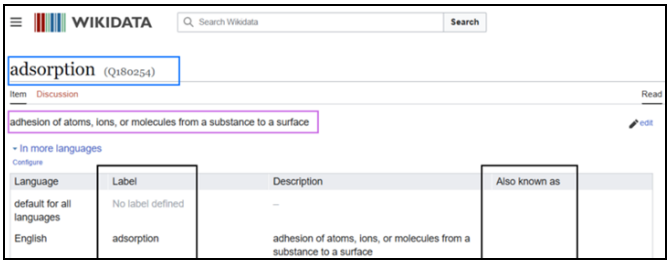


Figure 3 : Étiquetage conceptuel

3) Identification et association de concepts dans des *modlets*. Le terme *adsorption* va devenir à la fois label et concept d'une première version de *modlet* (*cf.* Figure 4, le rond orange représente un concept tandis que les ronds clairs représentent les labels associés à ce concept) ;

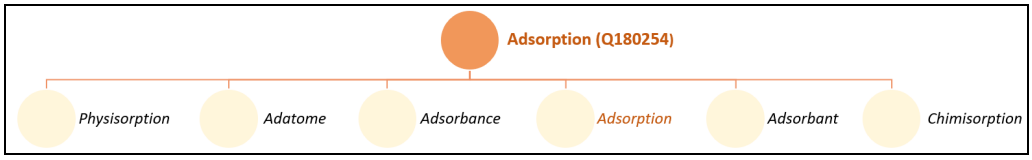


Figure 4 : Identification et association de concepts

4) Validation/correction des *modlets* avec les chercheurs. Ce *modlet*, construit à partir d'une publication, est validé ou corrigé par les chercheurs du PERL.

4.4. Enrichissement de modlets (tests, intégration de concepts, labels, relations, issus de ressources externes)

Les *modlets* construits à l'étape précédente sont désormais enrichis à l'aide de la ressource externe Wikidata. Des labels sont rajoutés (dans différentes langues) ainsi que des concepts (*cf.* Figure 5).

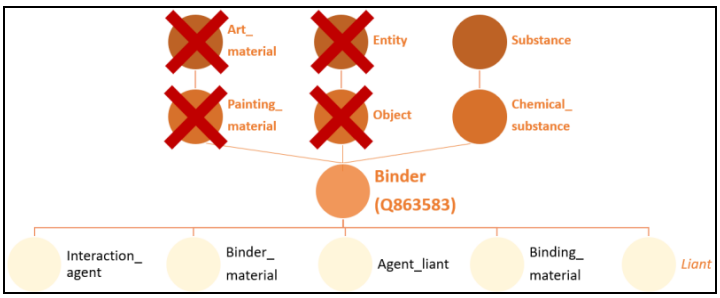


Figure 5 : Enrichissement de modlets

Dans ce second exemple, il faut noter que le terme *Liant* a été extrait d'une publication différente de celle qui a mis en exergue *Adsorption*. Ici, conformément au contexte, *Liant* est devenu un label rattaché au concept *Binder*. Les labels de couleur orange (ronds clairs) sont des termes chercheurs issus de publications scientifiques du PERL. Les labels en noir (ronds clairs) sont proposés par Wikidata. Les six concepts parents de *Binder* (ronds orange) sont également rajoutés sur la base de la ressource Wikidata. Ici, les chercheurs ont validé *Chemical_substance* et *Substance* mais ont supprimé quatre autres concepts sémantiquement éloignés du domaine.

4.5. Conception et tests de l'ontologie (intégration des modlets, expérimentations)

Les *modlets* sont intégrés pour constituer l'ontologie. À titre d'exemple (*cf.* Figure 6), nous prenons deux *modlets Adsorption* et *Electrophoresis* qui, dans la hiérarchie Wikidata, ont un même concept parent *Seperation_Process*. Ainsi, ceci nous permet de bâtir l'ontologie finale de manière itérative en assemblant les modlets deux à deux.

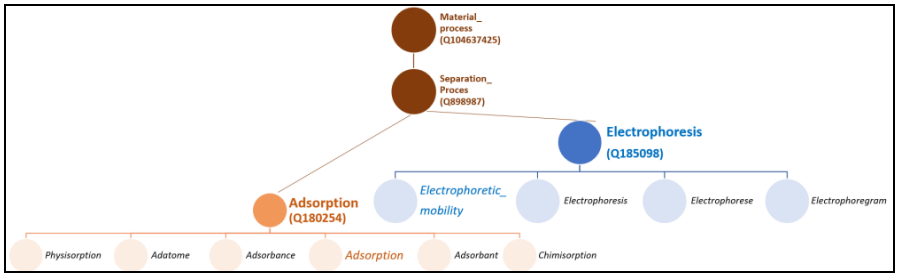


Figure 6 : Intégration de modlets

L'analyse terminologique issue du corpus scientifique du PERL met en évidence un total de 748 labels extraits des publications contenant effectivement 2335 occurrences. Le français représente 42% des labels, l'anglais domine avec 58%, conformément aux normes de production scientifique internationale et aux pratiques rédactionnelles des chercheurs du PERL. L'ontologie produite contient 350 concepts distincts (89 % des étiquettes provenant de Wikidata et 11 % des chercheurs du

PERL). La moyenne de 5,44 labels par concept constitue un indicateur déterminant : elle révèle l'existence d'une pluralité de formulations pour désigner un même objet scientifique, ce qui renforce la nécessité d'un travail de désambiguïsation. Les corrections, opérées par les cinq chercheurs mobilisés, ont permis de stabiliser les hiérarchies sémantiques et d'ajuster la base de connaissances en fonction des usages effectifs observés dans les publications du PERL.

La prépondérance des labels issus de Wikidata (89%) atteste par ailleurs de l'intégration massive de référentiels externes dans le processus de normalisation terminologique. L'ancrage dans Wikidata permet de stabiliser les identifiants conceptuels, de réduire les ambiguïtés lexicales et d'assurer l'interopérabilité des connaissances produites. À l'inverse, les labels spécifiquement issus du PERL (11%) témoignent de la présence d'un lexique local, situé, reflétant les pratiques discursives des chercheurs. Leur intégration dans l'ontologie garantit que le modèle ne se limite pas à un simple alignement sur un vocabulaire global, mais incorpore la dimension de la connaissance propre au terrain. Dans cette perspective, l'ontologie produite ne constitue pas uniquement un outil de formalisation, mais un dispositif de médiation visant à faire converger des ressources lexicales multiples vers une structure commune, mobilisable autant par les chercheurs que par les communicants.

Cette opération, bien que construite avec les experts du domaine, soulève cependant une question centrale : dans quelle mesure une base de connaissances patiemment construite, par un travail minutieux de tri, de confrontation et de validation experte, peut-elle pleinement déployer son potentiel tant qu'elle n'a pas été mise à l'épreuve d'usages réels ?

5. Preuve de concept

5.1. Le prototype CommTools

L'outil développé intitulé « *CommTools* » constitue la preuve de concept de notre méthode ETHCOMOD. Ce prototype met à l'épreuve la solidité du modèle en le confrontant à un usage réel dans le contexte de TotalEnergies. L'application permet de vérifier la robustesse sémantique de l'ontologie (cohérence, traçabilité, interopérabilité) tout en évaluant sa valeur médiatrice, c'est-à-dire sa capacité à favoriser le passage des savoirs scientifiques vers des formats communicationnels fiables.

L'architecture de *CommTools* a été pensée pour répondre à trois finalités principales :

(1) **Pour les chercheurs de l'entreprise**, elle constitue un espace de valorisation et de visibilité. Les publications indexées y apparaissent associées à des annotations conceptuelles issues de l'ontologie produite par ETHCOMOD, garantissant la fidélité des termes et la cohérence terminologique ;

(2) **Pour les communicants**, l'application représente un levier d'appropriation et de médiation. Elle leur donne accès à des contenus validés, contextualisés et non confidentiels, tout en offrant un niveau de lecture adapté à leurs

besoins de communication. Elle leur permet ainsi de construire des narratifs alignés sur la rigueur scientifique, tout en restant accessibles à des publics variés ;

(3) **Pour l'entreprise**, *CommTools* renforce la crédibilité des discours institutionnels autour de la R&D et participe à la construction d'une image d'innovation responsable et transparente.

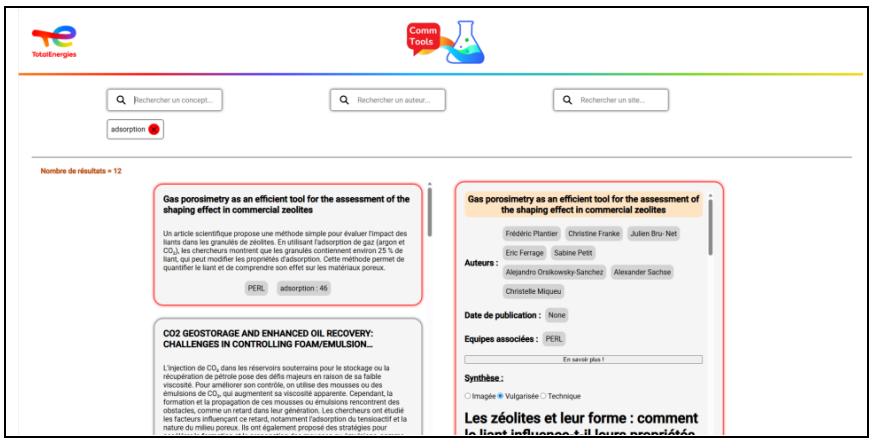


Figure 7 – Interface principale de l'outil CommTools

Cette première interface, visible dans la Figure 7, constitue le point d'entrée dans notre outil *CommTools*. Elle permet de rechercher un concept, un auteur ou un terme, grâce à un système d'*auto-complétion* connecté à l'index stocké dans une base GraphDB[1]. Cette base de connaissances RDF/OWL contient les annotations du corpus scientifique en lien avec l'ontologie métier construite par notre approche ETHCOMOD. Le moteur de recherche repose sur des requêtes SPARQL exécutées via une API Flask, qui interroge la base et retourne les résultats au format JSON. Les documents correspondants au résultat de la requête sont présentés sous forme de cartes enrichies (lorsque l'utilisateur sélectionne une carte celle-ci est entourée en rouge) : chaque carte affiche le titre, les auteurs, les équipes associées, et un résumé vulgarisé correspondant à des articles scientifiques du corpus répondant au critère de recherche de l'utilisateur de *CommTools*. Dans l'exemple illustré dans la Figure 7, l'utilisateur cherche des publications relatives au concept « *adsorption* » et 12 références de documents lui sont restitués. Évidemment, notre outil reconnaît automatiquement toutes les occurrences du concept « *adsorption* » dans le corpus, y compris lorsque le terme exact n'est pas utilisé (ex., si l'article emploie les expressions : *adsorbent*, *adsorption capacity* ou *surface interaction*).

Lorsque l'utilisateur souhaite disposer de plus de détails sur une publication renvoyée lors d'une recherche, une nouvelle interface permet de présenter plus

[1] https://graphdb.ontotext.com

d'informations sur l'article en lui-même comme on peut le voir dans la Figure 8. Sur la gauche on retrouve l'article en PDF, sur la droite un résumé vulgarisé et en bas un graphe interactif qui permet de naviguer au sein de l'index pour lequel on retrouve notamment les concepts, les auteurs ainsi que des relations associées. En sélectionnant par exemple le nœud « *adsorbants* » dans ce graphe, toutes les occurrences associées au concept sont mises en évidence dans le résumé ainsi que dans le PDF de l'article, les labels liés (*synonymes, termes équivalents, dérivés scientifiques*) s'affichent également, et la définition issue de Wikidata[2] ou de notre ontologie métier issue de TotalEnergies apparaît dans la zone d'information. Cette navigation au sein de l'index permet à l'utilisateur de visualiser graphiquement comment un terme ou un concept s'insère dans la logique scientifique de l'article (*concepts parents, sous-concepts, relations méthodologiques ou phénoménologiques*).

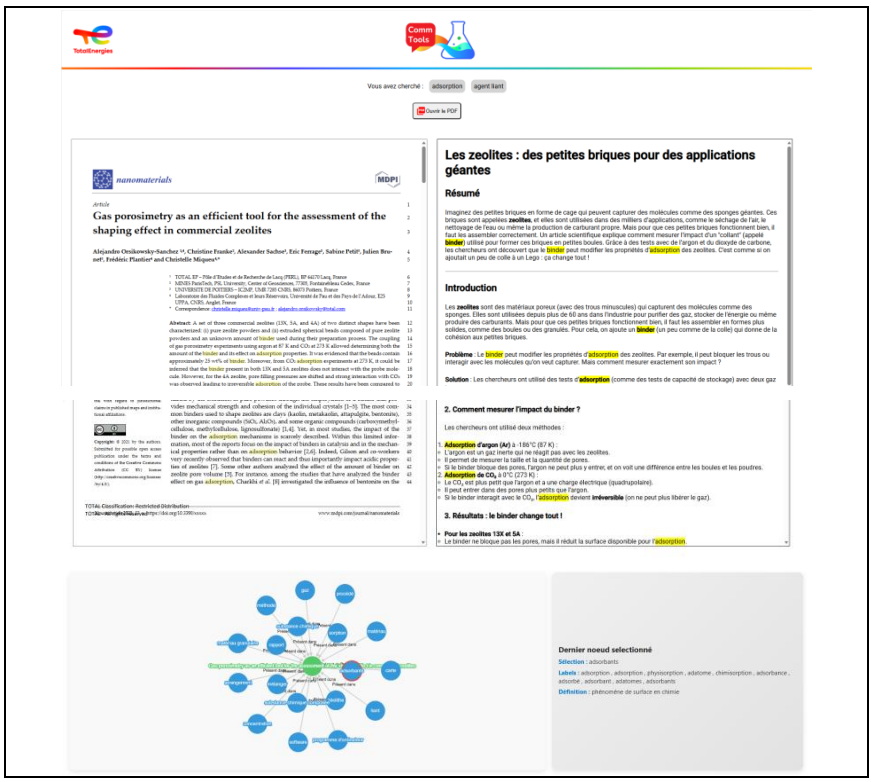


Figure 8 – Interface qui présente les informations au sujet d'un document

[2] https://www.wikidata.org

5.2. Évaluation avec les chercheurs et les communicants

Une expérimentation a été conduite pour évaluer les résultats générés par les interfaces de notre application *CommTools*, et en particulier la production par l'outil de textes vulgarisés et contextualisés via différents types de prompts prédéfinis. Cette démarche d'évaluation s'est appuyée sur les travaux de Kassogué et al. (2019) qui proposent une grille analytique structurée pour l'examen critique de productions scientifiques. L'enjeu de cette évaluation ne porte donc pas sur les prompts eux-mêmes en tant qu'instructions d'intelligence artificielle, mais plutôt sur la qualité, la fidélité et la lisibilité des textes générés, considérés comme objets de médiation entre la sphère scientifique et la sphère communicationnelle. Autrement dit, il s'agissait d'observer dans quelle mesure la génération assistée par ontologie, matérialisée dans *CommTools*, favorise une traduction intelligible, rigoureuse et légitime des savoirs scientifiques, apte à être mobilisée par les communicants dans leurs actions de valorisation de la R&D.

Pour chaque publication, trois versions de résumés vulgarisés sont soumises à évaluation (voir Darricades, 2026) : la première issue d'une génération automatique sans intégration ontologique (noté prompt A) ; la deuxième produite à partir des concepts proches de premier niveau de l'ontologie et leurs labels (noté prompt B) ; et la troisième (notée prompt C) qui exploite l'ensemble des concepts généralisants de l'ontologie et de leurs labels, offrant un niveau de contextualisation supérieur. Ces trois versions de résumés vulgarisés constituent des versions expérimentales destinées à comparer la qualité des prompts et à identifier celui produisant la traduction la plus satisfaisante. Le type de prompt retenu à l'issue de cette phase d'évaluation sera ensuite intégré de manière pérenne dans les interfaces concernées (cf., Figure 7 et Figure 8), afin d'optimiser la génération automatique de textes vulgarisés et contextualisés.

Chaque participant a évalué un ensemble de productions automatiques à l'aide d'une grille d'analyse construite sur la base de mesures quantitatives et d'observations qualitatives qui porte sur la fidélité scientifique (*présence et exactitude des notions centrales, rigueur du propos*), la clarté et la vulgarisation (*lisibilité, accessibilité du vocabulaire, compréhension pour un non-spécialiste*), la cohérence terminologique (*usage stable et pertinent des concepts issus de l'ontologie*), la capacité de médiation (*facilitation du dialogue chercheurs-communicants et communicants-grand public*), ainsi que la perception identitaire (*reconnaissance de soi dans le texte pour les chercheurs, exploitabilité pour les communicants*). Des espaces de commentaires libres permettent aux participants de motiver leurs choix, d'exprimer leurs perceptions et de proposer des ajustements.

Pour les chercheurs, l'évaluation du prompt A, généré sans ontologie, met en évidence une faiblesse notable dans la restitution des notions centrales : les chercheurs le jugent « dilué », imprécis et trop généraliste. La rigueur moyenne n'atteint que 3,2/5, et six chercheurs sur sept relèvent des erreurs ou omissions majeures. Ce résultat confirme que, sans structure sémantique sous-jacente, la génération automatique peine à maintenir la cohérence interne du discours scientifique et à restituer la hiérarchie des concepts. Le prompt B, intégrant des

concepts et labels proches issus de l'ontologie, marque une nette amélioration. La rigueur monte à 4/5, la fidélité scientifique est jugée satisfaisante par la majorité des chercheurs, et les commentaires soulignent une meilleure correspondance entre le texte généré et la logique méthodologique de la publication. L'ontologie joue donc ici un rôle de cadre de stabilisation du sens : elle aligne le vocabulaire sur la logique disciplinaire des chercheurs, tout en permettant une vulgarisation plus contrôlée. Le prompt C, enrichi par tous les concepts généralisants, obtient une évaluation globalement positive (rigueur 4/5, clarté 4,1/5), mais suscite des retours plus ambivalents : s'il est jugé plus fluide et accessible, il est aussi perçu comme moins précis scientifiquement. Certains chercheurs y voient un risque de simplification. Ces résultats montrent que l'intégration ontologique renforce la confiance des chercheurs envers le dispositif, tout en soulignant que la vulgarisation doit rester encadrée pour éviter la perte de profondeur scientifique.

Du côté des communicants, l'évaluation du prompt A, bien que jugé clair (4,4/5), est perçu comme descriptif et peu utile en communication stratégique, c'est-à-dire compréhensible mais sans profondeur. Le prompt B, plus détaillé, est salué pour sa rigueur mais jugé « dense », ce qui le rend moins exploitable pour des contenus de vulgarisation. Le prompt C s'impose en revanche comme la version la plus performante : il obtient les meilleures moyennes sur l'ensemble des indicateurs, 5/5 pour l'accessibilité, 4,5/5 pour la dynamique, 5/5 pour l'attractivité, et fait l'objet d'un consensus sur son intérêt communicationnel. Les communicants soulignent que le prompt C facilite l'appropriation rapide du sujet, notamment grâce à une contextualisation fluide et une hiérarchisation claire des idées, rendues possibles par l'appui ontologique. Sur le plan de l'usage, *CommTools* est perçu comme un outil de médiation assistée : il permet aux communicants de comprendre plus aisément des sujets techniques et d'en extraire les éléments clés pour la valorisation interne ou externe. Les communicants insistent également sur la nécessité d'une supervision humaine, rappelant que la médiation ne peut être automatisée sans une validation épistémique. Ce positionnement confirme leur compréhension du dispositif non comme une substitution, mais comme un amplificateur de la compréhension et de la circulation des savoirs.

La lecture croisée des deux panels (chercheurs d'un côté et communicants de l'autre) révèle une complémentarité forte mais différenciée entre les attentes des chercheurs et celles des communicants. Les chercheurs valorisent la fidélité scientifique et la cohérence terminologique, portées par le prompt B, tandis que les communicants privilégient la clarté, la fluidité et l'adaptabilité éditoriale, incarnées dans le prompt C. Ainsi, le prompt B est considéré par les chercheurs comme le plus fiable scientifiquement, car il restitue les concepts clés sans les simplifier à l'excès, tout en maintenant la structure logique des publications. De leur côté, les communicants identifient le prompt C comme le plus exploitable, en raison de sa capacité à rendre les thématiques scientifiques intelligibles, contextualisées et communicables, tout en conservant un niveau de rigueur satisfaisant grâce à l'appui ontologique.

6. Conclusion

La méthode ETHCOMOD démontre qu'une ontologie peut dépasser sa fonction technique de structuration de données pour devenir un véritable dispositif de médiation. En intégrant des ressources externes comme Wikidata et en impliquant activement les communautés métiers, cette approche permet de stabiliser un lexique commun tout en respectant les spécificités terminologiques locales. L'expérimentation menée avec le prototype *CommTools* confirme l'efficacité de cette démarche. L'évaluation croisée a révélé : que (1) le recours à l'ontologie garantit la fidélité scientifique indispensable aux chercheurs, et que (2) l'enrichissement sémantique (notamment via le prompt qui intègre des concepts généralisants) offre aux communicants la clarté et l'attractivité nécessaires à la valorisation des savoirs.

En conclusion, ETHCOMOD offre un cadre rigoureux pour assurer la circulation des connaissances dans un environnement industriel. Les perspectives de ce travail résident désormais dans la mise à l'épreuve de ces bases de connaissances au travers de multiples usages variés réels et quotidiens, afin de valider pleinement leur potentiel de transformation des pratiques communicationnelles en entreprise, et ainsi améliorer la communication interne et externe des organisations.

Ces résultats ouvrent alors une réflexion plus large sur la capacité des dispositifs ontologiques à reconfigurer durablement les modalités de circulation et de partage des savoirs au sein des organisations, interrogeant notamment la mesure dans laquelle l'intégration de référentiels mondiaux comme Wikidata dans des systèmes d'information locaux peut transformer durablement les pratiques de partage des savoirs, en faisant de l'ontologie non plus un simple dépôt de données, mais un véritable agent de confiance entre experts scientifiques et communicants.

Bibliographie

Aussenac-Gilles N., Despres S., Szulman S. (2008). The terminae method and platform for ontology engineering from texts. Paul Buitelaar and Philipp Cimiano, editors, Bridging the Gap between Text and Knowledge, pp 199-223

Darbellay, F. (2012). La circulation des savoirs. *Interdisciplinariété, concepts nomades, analogies, métaphores. Bern: Peter Lang*.

Darricades C. (2026). *Les contributions info-communicationnelles de la Recherche-Développement (R&D) à la traduction d'une mutation industrielle : le cas de TotalEnergies.* 2026, Thèse en Sciences de l'Information et de la Communication, Université de Lille.

Despres, S. (2014, May). Construction d'une ontologie modulaire pour l'univers de la cuisine numérique. In *IC-25èmes Journées francophones d'Ingénierie des Connaissances* (pp. 27-38).

El Mezouari, S., Lotfi, M., & Bouthir, Y. (2013). L'importance de la communication interne dans les entreprises: Cas d'une entreprise industrielle marocaine. *Revue Economie & Kapital*, (4).

Fernández-López, M., Gómez-Pérez, A. & Juristo Juzgado N. (1997). METHONTOLOGY: From Ontological Art Towards Ontological Engineering. *In AAAI-97 Spring Symposium Series*, 24-26 March 1997, Stanford University, EEUU.

Giroux, N., & Marroquin, L. (2005). L'approche narrative des organisations. *Revue française de gestion*, *159*(6), 15-42.

Grüninger, M., & Fox M.S., (1995). Methodology for the design and evaluation of ontologies. *In Proc. IJCAI'95, Workshop on Basic Ontological Issues in Knowledge Sharing*.

Hollink, L., Bedjeti, A., Van Harmelen, M., & Elliott, D. (2016, May). A corpus of images and text in online news. *In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 1377-1382).

Kassogué, B., Kassogué, P. T., & Dolo, S. (2019). Méthode de Recherche: Grille d'Analyse d'un Travail Scientifique. *International journal of Scientific and Research*, *10*(1), 1399-1402.

Lanciano-Morandat, C., Rolle, P., & Bouffartigue, P. S. (2019). Le travail de recherche. Production de savoirs et pratiques scientifiques et techniques. CNRS éditions.

Latour, B., & Biezunski, M. (2005). La science en action : introduction à la sociologie des sciences.

Peroni, S. (2016). SAMOD: an agile methodology for the development of ontologies. *In Proceedings of the 13th OWL: Experiences and Directions Workshop and 5th OWL reasoner evaluation workshop (OWLED-ORE 2016)* (pp. 1-14).

Rawsthorne, H. M., Abadie, N., Kergosien, E., Duchêne, C., & Saux, E. (2022, June). ATLANTIS: Une ontologie pour représenter les Instructions nautiques. *In IC 2022, 33es journées francophones d'Ingénierie des connaissances* (pp. 154-163).

Soegaard M. (2008). *Interaction Styles,* http://www.interaction-design.org/encyclopedia/interaction_styles.html

Staab, S., Studer, R., Schnurr, H. P., & Sure, Y. (2001). Knowledge processes and ontologies. *IEEE Intelligent systems*, *16*(1), 26-34.

Suàrez-Figuero M.C., Gòmez-Perez A., Motta E., Gangemi A. (2012). The NeOn Methodology for Ontology Engineering, Springer, pp 9-34

Swartout, B., Patil, R., Knight, K., & Russ, T. (1996, November). Toward distributed use of large-scale ontologies. In *Proc. of the Tenth Workshop on Knowledge Acquisition for Knowledge-Based Systems* (Vol. 138, No. 148, p. 25).

Vrandečić, D. (2012, April). Wikidata: A new platform for collaborative data collection. *In Proceedings of the 21st international conference on world wide web* (pp. 1063-1064).

Détection de motifs dans des Graphes temporels : Une Approche par la Logique Floue appliquée aux noms de domaines du Web

Marwa Alali[1,2], Arnaud Castelltort, Sebastian Cesario[2], Marzieh Derakhshannia[3], Anne Laurent[1]

1. LIRMM, Univ Montpellier, CNRS, Montpellier, France
marwa.alali@lirmm.fr

2. BFOREAI, Montpellier, France
https://bfore.ai

3. University of Toulouse, CNRS, IRIT, Toulouse, France

RESUME. *Les graphes de connaissances permettent la représentation structurée des données dans de nombreux domaines, mais l'intégration de la dimension temporelle reste peu explorée. Bien que les graphes de connaissances temporels incorporant des données historiques permettent un raisonnement sémantique plus riche, l'efficacité comparative des différentes approches d'intégration pour les réseaux de relations de domaines n'a pas été systématiquement évaluée. Cet article présente une analyse complète de trois méthodologies d'intégration de graphes pour la détection de faux domaines Web. Nous proposons une approche basée sur l'utilisation de la logique floue. Un système d'inférence de Mamdani à 27 règles a été créé, intégrant une expression de confiance interprétable. L'évaluation expérimentale montre que l'approche basée sur la logique floue réalise une synthèse d'objectifs complémentaires par rapport à l'approche standard. De nouvelles relations inter-temporelles sont découvertes grâce à cette représentation. Les scores de confiance interprétables du système d'inférence floue fournissent une quantification exploitable du risque, répondant à une limitation critique des approches conventionnelles d'intégration de graphes qui manquent de quantification de l'incertitude et d'explicabilité.*

MOTS-CLÉS : *graphes de connaissances, logique floue, intégration de graphes, analyse des relations de domaines, système d'inférence de Mamdani*

1. Introduction

La prolifération des dispositifs IoT, cloud computing et des réseaux 5G a exponentiellement augmenté le problème des URL et des noms de domaines sur le Web et, parallèlement, le nombre de faux domaines comme par exemple lnforsid.fr au lieu de inforsid.fr (Chen *et al.*, 2024; Velasco and Rodriguez, 2017). Les graphes de relations de domaines sont devenus des outils fondamentaux pour retrouver les faux domaines et comportements associés (Manadhata *et al.*, 2014). Ces graphes exploitent le principe que les domaines partageant une infrastructure d'hébergement reflètent souvent une propriété ou une intention opérationnelle commune (Khalil *et al.*, 2016; Khalil *et al.*, 2018; Nabeel *et al.*, 2020). Par exemple, les attaquants peuvent exploiter plusieurs sites web frauduleux qui vendent ostensiblement des jeux vidéo populaires. Bien que ces sites semblent distincts, ils partagent souvent une infrastructure sous-jacente telle que des systèmes de traitement des paiements, des adresses e-mail de contact ou des serveurs d'hébergement. En découvrant et en analysant ces similitudes infrastructurelles latentes, les systèmes peuvent identifier et attribuer des réseaux entiers d'activités frauduleuses coordonnées plutôt que de traiter chaque site Web de manière isolée (Xiao *et al.*, 2020). Les systèmes modernes emploient de plus en plus des graphes de provenance encodant les dépendances causales entre les entités pour la détection de menaces au sein des systèmes informatiques et des réseaux d'entreprise, démontrant leur efficacité dans la détection d'anomalies et l'analyse des causes profondes (Zipperle *et al.*, 2022).

Nous avons implémenté une solution, inspirée de Nabeel et Issa (Khalil *et al.*, 2018; Nabeel *et al.*, 2020), qui ingère quotidiennement de 500 000 à 5 millions de domaines Web nouvellement enregistrés (NRDs). Notre système construit des graphes de résolution de domaines sous forme de graphes bipartis non orientés $G(D, I, E)$, où D représente les domaines, I représente les adresses IP, et les arêtes E connectent les domaines à leurs IPs résolues. À partir de cette structure, nous dérivons des graphes de domaines $DG(D, E_{DG})$ où les arêtes E_{DG} reflètent les relations inférées via une infrastructure partagée. Chaque construction de graphe produit un instantané temporel isolé, limitant la visibilité complète des menaces et la découverte d'associations cachées à travers les périodes temporelles. Deux stratégies d'intégration conventionnelles présentent des limitations fondamentales. L'union simple ($G_{\text{union}} = \bigcup_{i=1}^{n} G_i$) préserve toutes les arêtes originales mais ne peut pas inférer les relations inter-graphes implicites. La reconstruction fusionne les données initiales et réapplique les algorithmes de détection, découvrant potentiellement des motifs cachés mais compromettant la provenance et écartant les arêtes originales, réduisant ainsi l'interprétabilité et la traçabilité. Ces limitations nécessitent une méthodologie d'intégration capable de préserver simultanément les relations existantes tout en révélant de nouvelles connexions inter-temporelles.

Le système d'intégration basé sur la logique floue proposé répond à ces défis en transformant les conditions d'analyse d'infrastructure de l'algorithme original en un mécanisme d'inférence floue de type Mamdani. Le système emploie des variables linguistiques pour quantifier la force de la relation sous incertitude, utilisant 27 règles définies par des experts pour évaluer l'appariement des couples de domaines. Cette

approche produit des scores de confiance interprétables pour les arêtes potentielles, maintenant l'intégrité structurelle tout en facilitant la découverte de comportements précédemment non observés.

Contributions :

1) Un système d'inférence floue de Mamdani à 27 règles avec cinq caractéristiques pour l'analyse des relations de domaines.

2) Un cadre d'intégration de N-graphes évolutif qui détecte des relations inter-temporelles invisibles dans les graphes grâce à l'inférence basée sur la logique floue.

3) L'optimisation du pré-filtrage par IP partagée atteignant une accélération de 25 600× tout en préservant la qualité de classification, permettant un déploiement pratique dans des environnements opérationnels.

4) La validation expérimentale sur des données réelles (10 graphes, 27 580 domaines, 1,2M+ arêtes) avec analyse comparative tripartite complète révélant les forces complémentaires des différentes approches d'intégration.

La suite de cet article est organisée comme suit. La Section 2 passe en revue les travaux connexes et identifie les limitations des approches existantes. La Section 3 détaille le cadre de logique floue proposé. La Section 4 présente les résultats expérimentaux et l'évaluation comparative. La Section 5 conclut l'étude.

2. Contexte et travaux connexes

L'adoption des graphes de connaissances a considérablement augmenté ces dernières années, s'étendant aux applications industrielles en raison de leurs avantages par rapport aux représentations de données traditionnelles (Dibowski, 2024). Dans de nombreuses applications réelles, les graphes de connaissances intègrent des indicateurs hétérogènes tels que les enregistrements DNS, les données WHOIS, les journaux réseau et les rapports de logiciels malveillants dans des structures unifiées qui prennent en charge l'analyse de corrélation et le raisonnement sur les menaces (Papoutsoglou *et al.*, 2024; Gao *et al.*, 2022).

Falcarin et al. (Falcarin and Dainese, 2024) lient les données de référentiels publics (par exemple, NVD, CWE, MITRE ATT&CK, CAPEC) pour modéliser les relations entre les données, tandis qu'Alharbi et al. (Alharbi *et al.*, 2025) emploient l'extraction d'entités basée sur le NLP et le raisonnement logique pour construire des graphes de connaissances à partir de jeux de données de vulnérabilités. Bien qu'efficaces pour consolider des sources statiques, ces méthodes se concentrent sur la découverte de relations intra-ensembles de données et ne traitent pas de l'intégration d'instantanés temporels de graphes générés par des systèmes opérationnels, l'écart ciblé dans ce travail.

La construction unifiée de graphes de connaissances fait face à des défis persistants incluant l'hétérogénéité des schémas, les identifiants d'entités incohérents, les indicateurs qui se chevauchent et les assertions contradictoires (Papoutsoglou *et al.*, 2024; Ho-

fer *et al.*, 2023). Les pipelines d'intégration naïfs produisent des entités de faible qualité, des relations fallacieuses et des difficultés de raisonnement (Alharbi *et al.*, 2025), nécessitant des approches fondées préservant l'intégrité sémantique et l'utilité opérationnelle. Les méthodes d'enchâssement de graphes de connaissances (Guo *et al.*, 2023) permettent la prédiction de liens sur des graphes statiques mais manquent d'intégration temporelle et de suivi de provenance, limitations directement abordées dans ce travail. Les algorithmes de base de relations de domaines (Khalil *et al.*, 2018; Nabeel *et al.*, 2020) emploient un seuillage binaire traitant tous les couples qualifiés uniformément, ne parvenant pas à capturer les nuances de force de partage d'infrastructure.

La logique floue a été de plus en plus appliquée dans de tels systèmes pour aborder l'incertitude et l'imprécision à travers son utilisation dans la détection de menaces, l'évaluation des risques et des anomalies, l'aide à la décision, et les modèles hybrides combinant le raisonnement flou avec l'IA ou les méthodes statistiques. La logique floue gère l'incertitude, l'ambiguïté et l'information incomplète à travers des fonctions d'appartenance et des variables linguistiques plutôt que des décisions binaires vrai/faux (Atitallah *et al.*, 2025; Dixit *et al.*, 2025). Des applications récentes démontrent l'efficacité dans la modélisation adaptative des menaces, la réduction des faux positifs et l'interprétabilité (Atitallah *et al.*, 2025; Dixit *et al.*, 2025). L'intégration de la logique floue avec des structures de graphes améliore la classification et la priorisation des menaces (Zhang, 2024), validant le raisonnement tenant compte de l'incertitude pour l'analyse de renseignements. Tan et al. (Tan *et al.*, 2026) démontrent dans le diagnostic de santé que les graphes de connaissances flous permettent l'intégration de données multi-modales en fusionnant des caractéristiques hétérogènes dans des bases de règles floues et en construisant des graphes avec des arêtes pondérées par appartenance.

À notre connaissance, les approches existantes dans ce domaine ont rarement recours à la logique floue pour préserver simultanément la provenance des arêtes, découvrir des relations inter-temporelles et fournir une quantification interprétable de la confiance, pourtant essentielle pour le renseignement opérationnel en cybersécurité. Notre cadre de logique floue aborde cela en modélisant la force de relation de manière continue à travers l'inférence multi-caractéristiques, en préservant la provenance complète et en générant des scores de confiance interprétables pour les connexions inter-temporelles.

3. Méthodologie

Le cadre proposé intègre plusieurs graphes temporels de relations de domaines en combinant la modélisation structurelle de graphes avec la logique floue pour capturer les caractéristiques topologiques et l'incertitude sémantique dans les données d'infrastructure. L'inférence floue transforme les caractéristiques d'hébergement imprécises en fonctions d'appartenance interprétables, permettant une évaluation graduée de la relation plutôt qu'une classification binaire. Cette approche hybride réalise une découverte de relations tenant compte de l'incertitude tout en préservant la provenance

des arêtes, facilitant une intégration temporelle robuste et abordant les limitations des méthodes déterministes. La Figure 1 fournit un résumé visuel de haut niveau du pipeline complet.

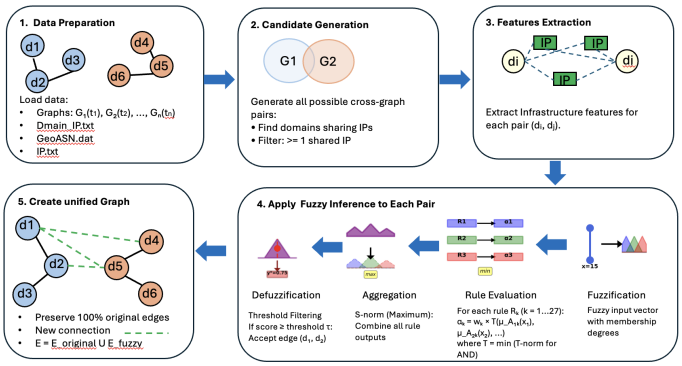


FIGURE 1 : Vue d'ensemble du cadre d'intégration multi-graphes basé sur la logique floue

3.1. *Préparation des données*

Le cadre traite des graphes d'entrée $G_i = (V_i, E_i)$ pour $i = 1, 2, \ldots, N$, où V_i désigne les nœuds (domaines) et E_i désigne les arêtes (relations). Les couples de nœuds et d'arêtes sont définis comme $V_{\text{all}} = \bigcup_{i=1}^{N} V_i$ et $E_{\text{original}} = \bigcup_{i=1}^{N} E_i$, respectivement.

Les jeux de données permettent l'extraction de caractéristiques et l'inférence : (1) corresponsdances domaine-vers-IP (`domain_IP.txt`) au format `domain ip1 ip2 ...`, (2) base de données ASN (`GeoASN.dat`) pour apparier les adresses IP , (3) métadonnées IP (`IP.txt`), et (4) listes d'IPs dédiées (hébergement mono-domaine) pour distinguer l'infrastructure exclusive de l'infrastructure partagée. Ces ressources permettent une extraction précise des relations et une interprétation sémantique pendant l'intégration.

3.2. *Génération de candidats*

Les arcs candidats inter-graphes sont identifiés entre des nœuds appartenant à des sous-graphes distincts. Pour chaque paire de graphes (G_i, G_j) telle que $i < j$, les candidats (u, v) vérifient $u \in V_i$, $v \in V_j$ et $(u, v) \notin E_{\text{original}}$. La contrainte stricte d'inter-graphe ($i \neq j$) impose implicitement un ordonnancement temporel : les arcs candidats ne sont inférés qu'entre des domaines issus d'instantanés temporels distincts, garantissant ainsi que l'inférence inter-graphe reflète un véritable chevauchement temporel d'infrastructure, plutôt qu'une simple cooccurrence au sein d'une même période.

Afin d'assurer la pertinence sémantique, les candidats sont ensuite filtrés de manière à ne conserver que les paires de domaines partageant au moins une adresse IP commune. Pour N graphes, ce processus examine $\binom{N}{2} = \frac{N(N-1)}{2}$ combinaisons de paires de graphes et, pour chaque paire, identifie les paires de domaines candidates exclusivement au moyen d'un index fondé sur les adresses IP, évitant ainsi une énumération exhaustive du produit cartésien.

Pour prendre en charge l'intégration à grande échelle, nous mettons en œuvre une génération optimisée des candidats, décrite dans l'algorithme 1 :

Algorithm 1 Génération optimisée des candidats

Construire l'index `ip_to_domains[ip][graph_id]` sur l'ensemble des graphes
Collecter E_{original} (tous les arcs existants, dans les deux directions)
for chaque paire de graphes (G_i, G_j) telle que $i < j$ **do**
    for chaque domaine $d_1 \in V_i$ apparaissant dans l'index IP **do**
        matches $\leftarrow \bigcup_{\text{ip} \in \text{IPs}(d_1)}$ `ip_to_domains[ip]`$[G_j] \cap V_j$
        for chaque $d_2 \in$ matches, avec $d_2 \neq d_1$ **do**
            if $(d_1, d_2) \notin E_{\text{original}}$ **then**
                Ajouter (d_1, d_2, G_i, G_j) à l'ensemble des candidats
            end if
        end for
    end for
end for
return candidats

3.3. *Extraction de caractéristiques*

Pour chaque couple candidat de domaines $(\text{domain}_i, \text{domain}_j)$, des caractéristiques basées sur l'infrastructure sont extraites en utilisant des jeux de données auxiliaires. Ces caractéristiques quantifient plusieurs dimensions des relations d'hébergement : l'étendue de l'utilisation d'adresses IP partagées entre les domaines, la diversité de l'infrastructure de routage réseau reflétée dans la distribution des numéros de systèmes autonomes (ASN), la présence de ressources d'hébergement dédiées versus partagées, la densité des domaines co-hébergés sur une infrastructure partagée, et les indicateurs de réputation de domaines provenant de listes curées. Ces caractéristiques sont dérivées de l'algorithme de relation binaire entre domaines proposé par Nabeel *et al.* (Nabeel *et al.*, 2020; Khalil *et al.*, 2018), dont les seuils rigides ont été reformulés sous forme de variables linguistiques floues afin de permettre une évaluation graduelle, plutôt que binaire, du partage d'infrastructure. Cette représentation multidimensionnelle permet au système flou d'évaluer la force de relation basée sur des motifs d'infrastructure plutôt que sur la connectivité binaire seule.

3.4. *Système d'Inférence floue*

La force de relation pour chaque couple candidat (d_i, d_j) est calculée via un système d'inférence floue de type Mamdani (Mamdani and Assilian, 1975; Zadeh, 1965) à travers quatre étapes :

3.4.1. *Fuzzification*

Les caractéristiques précises sont transformées en ensembles flous via des fonctions d'appartenance triangulaires $\mu_{\text{tri}}(x; a, b, c)$ et trapézoïdales $\mu_{\text{trap}}(x; a, b, c, d)$, où $a \leq b \leq c$ pour les fonctions triangulaires et $a \leq b \leq c \leq d$ pour les fonctions trapézoïdales, permettant des transitions douces entre les catégories linguistiques.

3.4.2. *Évaluation des règles*

Le système d'inférence floue de Mamdani emploie 27 règles SI-ALORS définies par des experts avec des poids d'importance associés ($w_k \in [0.7, 1.0]$), permettant un raisonnement interprétable sans nécessiter de données d'apprentissage étiquetées. Cette approche exploite l'agrégation conservative de la t-norme minimum pour réduire les faux positifs et incorpore des heuristiques pondérées pour refléter la force probante des caractéristiques basées sur l'infrastructure dans l'évaluation des relations de domaines. Chaque règle R_k suit :

$$R_k : \text{SI } x_1 \text{ est } A_{1k} \text{ ET } x_2 \text{ est } A_{2k} \text{ ET } \cdots \text{ ALORS } y \text{ est } B_k \quad [1]$$

La force d'activation α_k est calculée en utilisant la t-norme minimum :

$$\alpha_k = w_k \cdot \min(\mu_{A_{1k}}(x_1), \mu_{A_{2k}}(x_2), \ldots, \mu_{A_{nk}}(x_n)) \quad [2]$$

où w_k représente le poids de R_k, reflétant sa fiabilité relative. L'ensemble flou conséquent est obtenu via l'implication : $\mu^*_{B_k}(y) = \min(\alpha_k, \mu_{B_k}(y))$. Ce cadre permet de capturer des relations nuancées en combinant des règles spécifiques au domaine avec une inférence pondérée, fondée sur les évidences, dans un contexte d'incertitude.

Quelques exemples représentatifs de règles floues sont donnés ci-dessous :

– `IF ded_ip IS Many AND asn IS High AND shared_ip IS High` ⇒ `Very_Strong`

– `IF ded_ip IS None AND asn IS Medium AND shared_ip IS High` ⇒ `Moderate`

– `IF shared_ip IS Very_Low AND ded_ip IS None` ⇒ `Very_Weak`

3.4.3. *Agrégation et défuzzification*

Les ensembles conséquents modifiés s'agrègent via la s-norme maximum :

$$\mu_{\text{agg}}(y) = \max_{k=1}^{27} \min(\alpha_k, \mu_{B_k}(y)) \quad [3]$$

La défuzzification basée sur le centroïde produit des scores précis :

$$s = \frac{\int y \cdot \mu_{\text{agg}}(y)\, dy}{\int \mu_{\text{agg}}(y)\, dy} \quad [4]$$

Les arêtes avec des scores satisfaisant $s \geq \tau = 0.7$, correspondant aux catégories *Fort* et *Très Fort*, sont acceptées dans le graphe unifié. Ce seuil équilibre la découverte de relations significatives contre les connexions fallacieuses provenant d'une infrastructure partagée, tout en maintenant l'interprétabilité à travers un raisonnement linguistiquement fondé.

3.5. *Construction du graphe unifié*

Le graphe unifié $G_{\text{unified}} = (V, E)$ intègre tous les sous-graphes et les connexions inférées. L'ensemble de nœuds est $V = \bigcup_i V_i$, et l'ensemble d'arêtes est $E = E_{\text{original}} \cup E_{\text{fuzzy}}$, où $E_{\text{original}} = \bigcup_i E_i$ contient les arêtes originales et E_{fuzzy} inclut les arêtes inter-graphes inférées par logique floue.

Les nœuds sont annotés avec `graph_origins` indiquant les sous-graphes sources. Les arêtes portent : (1) `score` $\in [0, 1]$ (confiance floue), (2) `origins` (graphes sources ou étiquette `fuzzy`), et (3) `connecting_graphs` (par exemple, "$G_i \leftrightarrow G_j$") pour les arêtes inférées.

Cette approche préserve toutes les arêtes originales avec provenance complète tout en découvrant des motifs inter-temporels à travers l'inférence floue. La stratégie de pré-filtrage basée sur les IP réduit la complexité computationnelle sans perte de qualité, abordant les limitations des méthodes d'union (pas de découverte de relations implicites) et des approches de reconstruction (pas de traçabilité des arêtes).

4. Évaluation expérimentale

Cette section présente une évaluation concise du cadre d'intégration basé sur la logique floue proposé en utilisant dix graphes de domaines réels. Nous évaluons trois stratégies, Union, Reconstruction et Logique Floue, à travers la précision structurelle, la découverte de relations, l'interprétabilité et l'efficacité computationnelle.

4.1. *Configuration et jeu de données*

Les expériences ont été menées sur un MacBook Pro (M1/M2, ARM64, CPU 8-cœurs, 16 Go RAM) utilisant Python 3.13.3, NetworkX 3.6.1, et la base de données

MaxMind GeoIP ASN. Le jeu de données comprenait dix graphes de résolution de domaines d'octobre 2025, comprenant 27 580 nœuds uniques et 5,6 millions d'arêtes (voir Tableau 1).

Graphe	Nœuds	Arêtes
Graphe 1	4 420	6 133
Graphe 2	9 303	1 209 504
Graphe 3	4 611	5 906
Graphe 4	9 371	1 209 301
Graphe 5	4 507	6 208
Graphe 6	9 162	1 209 709
Graphe 7	4 391	5 918
Graphe 8	8 383	982 041
Graphe 9	4 366	6 128
Graphe 10	8 637	982 374
Total (unique)	**27 580**	**1,235,353**

TABLEAU 1 : Statistiques des graphes d'entrée (jeu de données d'octobre 2025)

4.2. *Résultats expérimentaux*

La Figure 2 et le Tableau 2 comparent les résultats. L'Union préserve la structure originale (27 580 nœuds, 1,23M arêtes). La Reconstruction génère 1,38M arêtes et introduit 261 nouveaux nœuds mais manque de provenance des arêtes. La Logique Floue conserve tous les nœuds originaux tout en ajoutant 644 arêtes inter-graphes à haute confiance avec interprétabilité.

Méthode	Nœuds	Arêtes	Nouvelles Arêtes
Reconstruction	27 841	1 385 124	—
Union	27 580	1 235 353	0
Logique Floue	27 580	1 235 997	644

TABLEAU 2 : Résultats structurels des méthodes d'intégration

La Figure 3 et le Tableau 3 montrent les chevauchements de nœuds et d'arêtes. Seulement 11 221 nœuds et 977 015 arêtes sont partagés à travers toutes les méthodes. La Reconstruction introduit 16 620 nœuds uniques et 407 465 arêtes. La Logique Floue et l'Union partagent le même ensemble de nœuds et 99,95% de similarité des arêtes, ne différant que par 644 arêtes supplémentaires inférées par logique floue.

La similarité de Jaccard (Tableau 4) confirme un accord structurel élevé entre Floue et Union (99,95%), avec une divergence significative de la Reconstruction (59,5%). Tous les arcs inférés par logique floue présentent des niveaux de confiance très resserrés, compris entre un minimum de 0,802 et un maximum de 0,806. Cette faible dispersion des scores constitue une conséquence mathématiquement attendue de

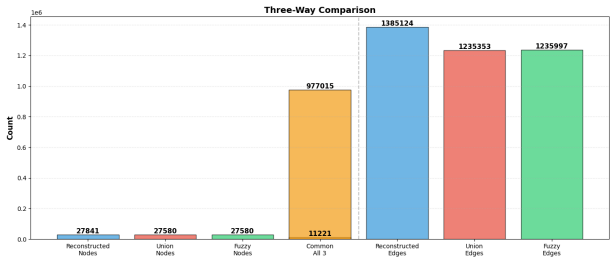


FIGURE 2 : Comparaison de la taille des graphes à travers les méthodes d'intégration

Chevauchement Nœuds	**Compte**	**Chevauchement Arêtes**	**Compte**
Les trois méthodes	11 221	Les trois méthodes	977 015
Union & Floue seulement	16 359	Union & Floue seulement	258 338
Reconstruction seulement	16 620	Reconstruction seulement	407 465

TABLEAU 3 : Chevauchement de nœuds et d'arêtes à travers les méthodes d'intégration

l'application d'une défuzzification par centroïde à des entrées à valeurs entières. En effet, des variables discrètes ne peuvent engendrer qu'un ensemble fini de combinaisons d'activation des règles, lequel se projette sur un nombre restreint de valeurs de sortie discrètes, plutôt que sur une distribution continue.

Dans notre jeu de données, la diversité des ASN apparaît comme le seul facteur discriminant. Plus précisément, les valeurs 5, 6 et 7–9 conduisent respectivement à des scores de $0{,}802$, $0{,}804$ et $0{,}806$. Ce comportement s'explique par le fait que chaque incrément entier le long de la pente croissante de la fonction d'appartenance « Medium » déplace le centroïde d'exactement $0{,}002$.

À l'inverse, l'utilisation de fonctions d'appartenance gaussiennes produit, même pour des entrées entières, des degrés d'appartenance irrationnels (transcendants). Les forces d'activation des règles deviennent alors quasi continues, ce qui rend le système particulièrement sensible au réglage des paramètres. De faibles variations des centres ou des variances peuvent ainsi suffire à faire passer les sorties en dessous du seuil d'acceptation, ce qui rend compte de l'absence totale de nouveaux arcs observée dans notre expérimentation fondée sur des fonctions gaussiennes.

La Figure 4 montre que les arêtes découvertes par logique floue sont concentrées entre des couples de graphes spécifiques, indiquant un chevauchement d'infrastructure temporel.

Comme le montre le Tableau 5, la Logique Floue a traité 113 107 couples candidats en 34,53 secondes, atteignant une accélération de $25\,600\times$ par rapport à la comparaison

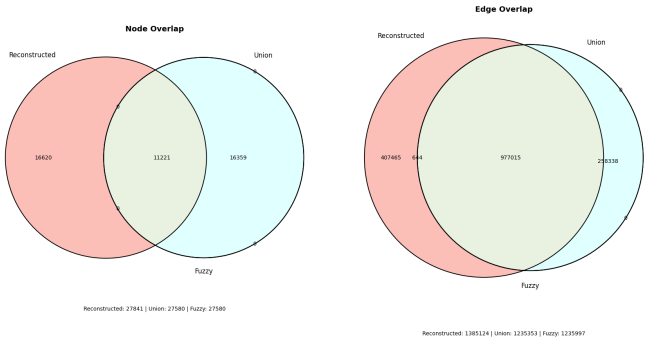


FIGURE 3 : Chevauchement de nœuds et d'arêtes

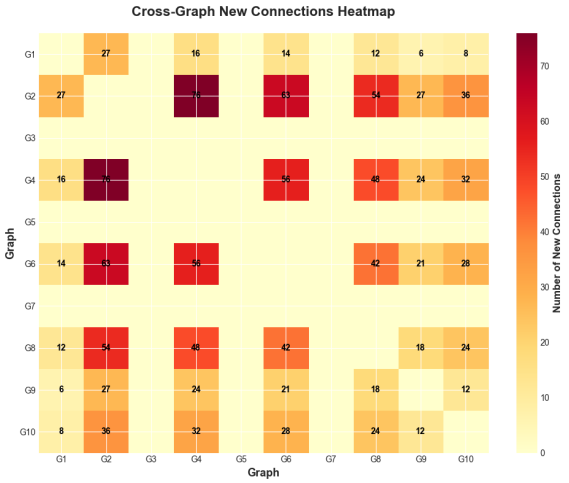


FIGURE 4 : Carte de chaleur des connexions inter-graphes inférées par logique floue

par couple exhaustive en exploitant le filtrage basé sur les IP. Elle conserve 100% de provenance et fournit des scores interprétables, contrairement à la Reconstruction.

Nous avons évalué l'impact du choix de fonction d'appartenance sur la performance d'intégration de graphes flous en comparant les modèles Gaussien et Triangulaire/Trapézoïdal sous des paramètres d'inférence identiques. Le Tableau 6 illustre cette comparaison. L'ajustement de paramètres a été appliqué au modèle Gaussien (centres et sigmas) pour aligner sa distribution de sortie avec celle du système triangulaire. Les deux modèles ont utilisé le même espace de caractéristiques d'entrée, univers

Comparaison	Arêtes Communes	Indice de Jaccard
Reconstruction ∩ Union	977 015	59,45%
Reconstruction ∩ Floue	977 659	59,49%
Union ∩ Floue	1 235 353	99,95%

TABLEAU 4 : Similarité de Jaccard par couples des ensembles d'arêtes

Métrique	Reconstruction	Union	Floue
Temps d'exécution	—	<1s	34,53s
Couples candidats évalués	—	0	113 107
Accélération	—	—	25 600×
Rétention de provenance	59%	100%	100%
Interprétabilité	Faible	Aucune	Élevée

TABLEAU 5 : Résumé de performance computationnelle

du discours et seuil de confiance ($\tau = 0.7$). Comme attendu dans l'inférence floue de Mamdani, toutes les 27 règles ont été activées dans les deux systèmes, chacune contribuant à des degrés variables d'influence basés sur l'appartenance d'entrée. Cependant, le système triangulaire/trapézoïdal a identifié 644 nouvelles arêtes inter-graphes, tandis que le système basé sur le Gaussien n'en a produit aucune. Ce résultat souligne la sensibilité supérieure des fonctions d'appartenance linéaires par morceaux dans la capture de relations subtiles basées sur l'infrastructure au sein de distributions d'entrée éparses.

Métrique	Triangulaire/Trapézoïdal	Gaussien
Total de nœuds	27 580	27 580
Arêtes originales	1 235 353	1 235 353
Nouvelles arêtes floues	644	0
Total d'arêtes finales	1 235 997	1 235 353
Candidats évalués	113 107	113 611
Temps d'exécution (s)	34,53	26,68

TABLEAU 6 : Comparaison des modèles de fonctions d'appartenance pour l'intégration de graphes flous

Ces résultats soutiennent l'adoption de fonctions d'appartenance Triangulaires/Trapézoïdales pour l'intégration de graphes de domaines, particulièrement lorsque des distinctions subtiles dans la similarité d'infrastructure doivent être capturées avec une inférence interprétable basée sur des règles.

4.3. *Discussion*

Les résultats expérimentaux démontrent que l'intégration basée sur la logique floue proposée réalise une synthèse d'objectifs complémentaires inatteignables par les méthodes conventionnelles. L'Union fournit une exécution instantanée et une préservation complète des arêtes mais ne découvre aucune relation inter-graphes, limitant la détection de motifs temporellement distribués. La Reconstruction élargit substantiellement la couverture (407K arêtes supplémentaires) mais sacrifie la provenance et l'interprétabilité. L'approche floue comble cette division, préservant toutes les arêtes originales tout en ajoutant 644 connexions interprétables avec scores de confiance, maintenant 99,95% de similarité structurelle avec l'Union. Cette augmentation sélective exploite les caractéristiques basées sur l'infrastructure (IPs partagées, diversité ASN) pour une inférence sémantiquement fondée.

L'évolutivité computationnelle représente une avancée critique. Le pré-filtrage basé sur les IP a réduit les couples candidats de 2,89 milliards à 113 107 (réduction de 99,996%), permettant une accélération de $25\ 600\times$ avec un temps d'exécution total de 34,53 secondes, confirmant la faisabilité pour un déploiement en temps réel de renseignement sur les menaces.

La sélection de fonction d'appartenance impacte significativement l'efficacité d'inférence. Sous des configurations identiques, le modèle Triangulaire/Trapézoïdal a découvert 644 nouvelles connexions tandis que le modèle Gaussien n'en a inféré aucune, malgré un coût computationnel quasi équivalent (différence de temps de 7,6%). Ce contraste frappant démontre que les fonctions linéaires par morceaux présentent une sensibilité supérieure aux motifs d'infrastructure subtils dans le raisonnement flou critique pour la sécurité, validant la conception soigneuse des fonctions d'appartenance comme essentielle pour les systèmes opérationnels.

Le cadre capture avec succès les relations de domaines inter-temporelles latentes avec stabilité structurelle, transparence sémantique et efficacité opérationnelle, représentant une avancée substantielle par rapport aux alternatives déterministes et basées sur la reconstruction pour les applications de reconnaissance de domaines.

5. Conclusion

Ce travail a introduit un cadre d'intégration basé sur la logique floue pour les graphes de connaissances qui améliore la détection des relations de domaines Web à travers les instantanés temporels. En appliquant un système d'inférence floue de Mamdani sur des caractéristiques dérivées de l'infrastructure, la méthode a découvert avec succès des arêtes inter-graphes à haute confiance tout en préservant 100% des connexions originales.

L'évaluation extensive sur des données réelles (octobre 2025, 10 graphes, 27 580 nœuds, 1,23'M arêtes) a démontré la praticité du cadre, atteignant une accélération de 25 600× via l'optimisation basée sur les IP, avec seulement 34,53 secondes de

temps d'exécution. L'analyse comparative a confirmé que l'intégration floue offre un terrain d'entente stratégique entre la préservation de base (Union) et la reconstruction exhaustive, équilibrant découverte, provenance, interprétabilité et évolutivité.

Les travaux futurs pourront explorer des modèles hybrides combinant des systèmes flous basés sur des règles et sur l'apprentissage, l'ajustement adaptatif d'appartenance et l'intégration de dynamiques temporelles pour améliorer davantage le raisonnement dans des paysages de menaces évolutifs.

Bibliographie

Alharbi H., Hur A., Alkahtani H., Ahmad H. F., « Enhancing cybersecurity through autonomous knowledge graph construction by integrating heterogeneous data sources », *PeerJ Computer Science*, vol. 11, p. e2768, 2025.

Atitallah S. B., Driss M., Boulila W., Koubaa A., « Securing Industrial IoT Environments : A Fuzzy Graph Attention Network for Robust Intrusion Detection », *IEEE Open Journal of the Computer Society*p. 1-12, 2025.

Chen H., Shen Z., Wang Y., Xu J., « Threat Detection Driven by Artificial Intelligence : Enhancing Cybersecurity with Machine Learning Algorithms », *World Journal of Innovation and Modern Technology*, vol. 7, n° 6, p. 58-70, 2024.

Dibowski H., « Full Traceability and Provenance for Knowledge Graphs », *Formal Ontology in Information Systems*, IOS Press, p. 223-237, 2024.

Dixit A., Bagde P., Hardiya P., « Advanced Cyber Security Using Fuzzy Logic », *Journal of Emerging Technologies and Innovative Research*, 2025. Article ID : JETIR2511119.

Falcarin P., Dainese F., « Building a Cybersecurity Knowledge Graph with CyberGraph », *Proceedings of the ACM/IEEE International Workshop on Engineering and Cybersecurity of Critical Systems and the IEEE/ACM International Workshop on Software Vulnerability*, p. 29-36, 2024.

Gao Y., Li X., Peng H., Fang B., Yu P. S., « HinCTI : A Cyber Threat Intelligence Modeling and Identification System Based on Heterogeneous Information Network », *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, n° 2, p. 708-722, 2022.

Guo Q., Liao Y., Li Z., Lin H., Liang S., « Convolutional Models with Multi-Feature Fusion for Effective Link Prediction in Knowledge Graph Embedding », *Entropy*, vol. 25, n° 10, p. 1472, 2023.

Hofer M., Obraczka D., Saeedi A., Köpcke H., Rahm E., « Construction of Knowledge Graphs : State and Challenges », *arXiv preprint arXiv :2302.11509*, 2023.

Khalil I. M., Guan B., Nabeel M., Yu T., « A Domain is Only as Good as Its Buddies : Detecting Stealthy Malicious Domains via Graph Inference », *Proceedings of the Eighth ACM Conference on Data and Application Security and Privacy*, p. 330-341, 2018.

Khalil I., Yu T., Guan B., « Discovering malicious domains through passive DNS data graph analysis », *Proceedings of the 11th ACM Asia Conference on Computer and Communications Security*, p. 663-674, 2016.

Mamdani E. H., Assilian S., « An experiment in linguistic synthesis with a fuzzy logic controller », *International Journal of Man-Machine Studies*, vol. 7, n° 1, p. 1-13, 1975.

Manadhata P., Yadav S., Rao P., Horne W., « Detecting malicious domains via graph inference », *Proceedings of the Workshop on Artificial Intelligence and Security*, p. 59-60, 2014.

Nabeel M., Khalil I. M., Guan B., Yu T., « Following Passive DNS Traces to Detect Stealthy Malicious Domains via Graph Inference », *ACM Transactions on Privacy and Security*, vol. 23, n° 4, p. 1-36, 2020.

Papoutsoglou M., Meditskos G., Bassiliades N., Kontopoulos E., Vrochidis S., « Mapping the Current Status of CTI Knowledge Graphs through a Bibliometric Analysis », *Proceedings of the 13th Hellenic Conference on Artificial Intelligence*, p. 1-6, 2024.

Tan N. H., Tuan T. M., Chuan P. M., Hoang N. D., Thanh L. Q., Son L. H., « FKG-MM : A Multi-Modal Fuzzy Knowledge Graph with Data Integration in Healthcare », *PLOS ONE*, vol. 21, n° 1, p. e0339864, 2026.

Velasco D., Rodriguez G., « Ontologies for Network Security and Future Challenges », *arXiv preprint arXiv :1704.02441*, 2017.

Xiao Q., Liu J., Wang Q., Jiang Z., Wang X., Yao Y., « Towards Network Anomaly Detection Using Graph Embedding », *International Conference on Computational Science*, Springer International Publishing, Cham, p. 156-169, 2020.

Zadeh L. A., « Fuzzy sets », *Information and Control*, vol. 8, n° 3, p. 338-353, 1965.

Zhang Q., « Optimizing Threat Intelligence Strategies for Cybersecurity Awareness Using MADM and Hybrid GraphNet-Bipolar Fuzzy Rough Sets », *International Journal of Advanced Computer Science and Applications*, 2024.

Zipperle M., Gottwalt F., Chang E., Dillon T., « Provenance-Based Intrusion Detection Systems : A Survey », *ACM Computing Surveys*, vol. 55, n° 7, p. 1-36, 2022.

Vers une Modélisation Générique et Eco-responsable de la Résolution d'Entités

Zhongwei MA [1,2], **Philippe ROOSE** [2], **Jiefu SONG** [3]

1. Technopôle Domolandes, Saint Geours de Maremne, France

2. Universite de Pau et des Pays de l'Adour, E2S UPPA, LIUPPA Anglet, France
zhongwei.ma@univ-pau.fr,philippe.roose@univ-pau.fr

3. Institut de Recherche en Informatique de Toulouse - Université Toulouse Capitole 31000, Toulouse, France
jiefu.song@ut-capitole.fr

RÉSUMÉ. *La généralisation des applications d'intelligence artificielle s'accompagne d'un besoin croissant en données volumineuses et hétérogènes. Dans ce cadre, la résolution d'entités (ER) constitue un moyen central pour rapprocher et combiner des informations issues de sources multiples, en vue de constituer des ensembles de données cohérents pour des usages spécifiques. Les travaux existants en ER proposent cependant un large éventail de méthodes. Ces méthodes sont souvent élaborées dans des cadres théoriques distincts et évaluées selon des critères propres à chaque approche. Les relations conceptuelles qui les relient restent rarement formulées de manière explicite. Dans cet article, nous proposons un modèle conceptuel dont l'objectif est d'identifier des notions communes à plusieurs lignes de recherche en ER et de les organiser dans une structure unifiée. Ce cadre vise à offrir une vision transversale des différentes approches. En parallèle, la question de la consommation énergétique demeure peu étudiée de façon structurée dans ce domaine, alors même qu'elle prend une importance croissante en systèmes informatiques. Nous intégrons donc explicitement cette dimension dans la modélisation du processus de l'ER, afin de la considérer conjointement aux indicateurs classiques de performance. Enfin, nous illustrons, par plusieurs cas représentatifs, que des méthodes d'ER existantes peuvent être décrites au moyen du cadre générique proposé, ce qui met en évidence sa capacité à couvrir des approches de nature variée.*

ABSTRACT.

MOTS-CLÉS : *résolution d'entité, énergie, modélisation conceptuelle*

KEYWORDS: *entity resolution, energy, conceptual modeling*

1. Introduction

Les récentes avancées en intelligence artificielle s'accompagnent d'un besoin croissant en données fiables et cohérentes, ce qui renforce le rôle de la résolution d'entités (ER). L'ER constitue une opération fondamentale de l'intégration des données, puisqu'elle vise à détecter et regrouper les descriptions qui désignent une même entité du monde réel (Christophides *et al.*, 2020). Cette tâche devient particulièrement complexe lorsque les données proviennent de sources multiples et indépendantes, dépourvues d'identifiants partagés. L'ER constitue alors une étape clé de la gestion et de la préparation des données, en détectant les divergences entre sources et en établissant des liens entre des enregistrements connexes, afin de fournir des données cohérentes aux traitements et aux applications en aval, par exemple aux applications d'IA.

Malgré son importance et une activité de recherche soutenue sur plusieurs décennies, l'ER se caractérise aujourd'hui par une grande hétérogénéité des approches. Les premières méthodes reposaient principalement sur des règles explicites et des mesures de similarité, formulées à partir d'hypothèses structurelles fortes (Gazzarri, Herschel, 2023 ; Nanayakkara, Christen, 2022). Ces hypothèses, comme l'homogénéité des formats ou la stabilité des valeurs d'attributs, sont rarement vérifiées dans des environnements réels, où les données sont souvent bruitées, hétérogènes ou ambiguës, ce qui conduit à de nombreuses variations dans les expressions des attributs (Binette, Steorts, 2022). L'émergence de l'apprentissage automatique a progressivement enrichi ce cadre, en introduisant des modèles capables de capturer des similarités textuelles complexes et des relations sémantiques plus fines. Les travaux récents s'orientent soit vers l'adaptation de modèles génériques, tels que les architectures de type Transformer, aux spécificités de l'ER (Gao *et al.*, 2023), soit vers la conception de chaînes de traitement fortement dépendantes d'algorithmes particuliers (Jabrane *et al.*, 2024). Cette accumulation de solutions hétérogènes complique toutefois la comparaison, la réutilisation et le choix éclairé des méthodes ER des utilisateurs.

Bien que les approches basées sur l'apprentissage automatique ont permis d'améliorer significativement les résultats de l'ER, généralement évalués à l'aide de la *précision*, du *rappel* et du *F1 score*, elles introduisent également des exigences accrues en ressources computationnelles, et donc par extension en énergie. Plusieurs études tentent d'estimer ces coûts à travers le temps d'exécution (Zhu *et al.*, 2025 ; Zeakis *et al.*, 2023). Néanmoins, le temps d'exécution ne capture qu'imparfaitement l'impact réel de ces méthodes, alors que l'étude de la consommation d'énergie des algorithmes d'ER fournit une mesure plus directe et plus pertinente. Cette dimension devient d'autant plus critique dans un contexte marqué par des engagements internationaux en faveur de la réduction de la consommation énergétique et des émissions de carbone, tels que l'Accord de Paris et les objectifs de développement durable des Nations Unies. À notre connaissance, la consommation d'énergie n'a toutefois jamais été considérée comme un critère d'évaluation central dans les travaux sur l'ER. Cette absence met en évidence un décalage entre les avancées techniques du domaine et les enjeux actuels de durabilité.

Compte tenu de ces défis, notre travail poursuit deux objectifs. Premièrement, nous visons à fournir une vue d'ensemble qui permette aux utilisateurs de sélectionner et d'appliquer le pipeline ER approprié. Deuxièmement, en formalisant l'énergie comme une dimension d'évaluation, nous introduisons une double perspective sur l'ER, la transformant d'une tâche uniquement centrée sur la performance en une tâche capable de représenter et de raisonner explicitement sur la relation entre la performance et la consommation d'énergie.

Pour atteindre ces objectifs, nous apportons deux contributions clés : (i) nous proposons une modélisation générale d'ER qui formalise l'énergie comme une dimension d'évaluation importante, permettant à l'ER d'être exprimée comme un processus à double perspective dans lequel la performance et l'énergie peuvent être représentées et analysées conjointement ; (ii) nous démontrons la généralité de cette modélisation à travers de multiples cas d'usages et sa faisabilité au moyen d'une implémentation modulaire.

Le reste de l'article est organisé comme suit. La section 2 passe en revue les approches existantes de l'ER et motive la nécessité d'une perspective d'exécution unifiée. La section 3.1 introduit nos concepts fondamentaux. La section 3.2 présente le cadre proposé et son intégration de l'énergie en tant que dimension d'évaluation. La section 4 illustre la faisabilité de notre cadre. Enfin, la section 5 conclut l'article.

2. État de l'art

Nous avons mené une analyse des revues Q1 et des publications des conférences CORE A/A* afin de caractériser le paysage actuel de l'ER. Nous avons sélectionné 29 articles sur la base de trois mots-clés : résolution d'entités (entity resolution), correspondance d'entités (entity matching) et couplage d'enregistrements (record linkage). Parmi ces articles, 25 ont été publiés entre 2023 et 2025, et nous avons inclus 4 travaux influents datant de 2018 afin de couvrir un large éventail de méthodologies. Selon ces articles, l'ER est devenu un domaine riche mais fragmenté, avec des techniques réparties sur quatre étapes différentes d'un pipeline opérationnel.

Prétraitement transforme les enregistrements bruts afin de les préparer à la comparaison et d'améliorer la qualité de leur représentation. Les opérations courantes (Gazzarri, Herschel, 2023) incluent la normalisation, la conversion de types et la standardisation des attributs. Bien que la plupart des travaux supposent implicitement des données déjà standardisées et cohérentes en type, cette condition préalable est rarement explicitée ou traitée de manière systématique.

Génération de candidats identifie des paires d'enregistrements tout en évitant les comparaisons quadratiques (Papadakis, Fisichella *et al.*, 2023 ; R. Wang, Zhang, 2024), ce qui réduit l'espace de recherche pour les étapes de rapprochement suivantes. Les enregistrements sont regroupés en blocs à l'aide de jetons partagés, de voisinages triés (Maciejewski *et al.*, 2025) ou de signatures basées sur le hachage (Nanayakkara, Christen, 2022), et seules les paires au sein d'un même bloc sont formées. Des va-

riantes préservant la confidentialité sécurisent les attributs à l'aide de filtres de Bloom ou d'encodages cryptographiques (Ziyad *et al*., 2025 ; Randall *et al*., 2024). Cette étape produit un ensemble réduit de paires candidates (Zhu *et al*., 2025) pour les traitements ultérieurs, correspondant à des enregistrements susceptibles de correspondre.

Comparaison par pairs calcule des scores de similarité pour les paires candidates. Elle s'appuie sur la sortie de la génération de candidats et constitue le cur de la phase de comparaison du pipeline d'ER. Les méthodes traditionnelles (Moretti, Shlomo, 2023) utilisent des fonctions de similarité sur des chaînes ou des valeurs numériques, tandis que les approches par graphes de similarité (Papadakis, Efthymiou *et al*., 2023) intègrent des informations relationnelles. Au-delà des règles de similarité conçues manuellement, l'apprentissage automatique introduit des modèles entraînant des classificateurs pour décider des correspondances (Wu *et al*., 2020 ; Jabrane *et al*., 2024). L'apprentissage profond utilise des réseaux de neurones (Mudgal *et al*., 2018), offrant des représentations contextuelles et une meilleure robustesse face à des données hétérogènes et bruitées. Cette direction est ensuite renforcée par des encodeurs de type Transformer pour des représentations sémantiques plus riches (Li *et al*., 2020 ; Gao *et al*., 2023 ; Liu, Shen, 2025 ; Ding *et al*., 2024 ; Wadhwa *et al*., 2024 ; R. Wang, Zhang, 2024 ; Arora, Dell, 2024 ; Dou *et al*., 2023), y compris des travaux visant à améliorer leur interprétabilité (Baraldi *et al*., 2023). Les modèles de langage de grande taille (LLM) réduisent récemment le besoin d'annotations grâce à des connaissances sémantiques pré-entraînées (Peeters *et al*., 2024 ; Zhu *et al*., 2025 ; Y. Wang *et al*., 2024 ; Zhang *et al*., 2025 ; T. Wang *et al*., 2024). La sortie de cette étape est un ensemble de décisions d'appariement par paires, qui sert de base aux étapes collectives ou à la consolidation finale des entités.

Comparaison collective exploite les dépendances globales pour affiner et coordonner les décisions d'appariement (Christophides *et al*., 2020). Les approches symboliques imposent une cohérence globale, soit par la transitivité des clusters, soit par des contraintes logiques (Bienvenu *et al*., 2023 ; Fagin *et al*., 2023). Les modèles neuronaux propagent l'information relationnelle entre entités, les réseaux de neurones sur graphes capturant directement les dépendances structurelles (Hu *et al*., 2025 ; Yao *et al*., 2022). Enfin, les systèmes hybrides combinent représentations apprises et contraintes symboliques pour produire des décisions plus robustes (Zhu *et al*., 2025). Contrairement aux approches pair à pair, les méthodes collectives fournissent des décisions de fusion globalement cohérentes.

Malgré des avancées importantes, la recherche en ER reste fragmentée avec des techniques et des pipelines hétérogènes. Les systèmes orientés processus cherchent à intégrer le prétraitement, le blocage, l'appariement et l'inférence collective dans des chaînes complètes et déterministes (Gazzarri, Herschel, 2023 ; Maciejewski *et al*., 2025 ; Papadakis *et al*., 2020). Ces pipelines sont assemblés de manière opérationnelle et leurs composants interagissent via des interfaces dépendantes de l'implémentation.

L'évaluation de ces processus se fait selon deux dimensions : la qualité des décisions d'appariement et le coût de calcul pour les produire. La plupart des travaux se concentrent sur des mesures de performance comme la précision, le rappel et le F1

score. Les évaluations orientées coût considèrent les budgets de comparaison (Maciejewski *et al*., 2025 ; Gazzarri, Herschel, 2023 ; Hu *et al*., 2025 ; Nanayakkara, Christen, 2022 ; R. Wang, Zhang, 2024 ; Papadakis, Fisichella *et al*., 2023) ou le temps d'exécution (Zeakis *et al*., 2023 ; Hu *et al*., 2025 ; Ranbaduge *et al*., 2024 ; Jabrane *et al*., 2024 ; Papadakis, Efthymiou *et al*., 2023 ; Arora, Dell, 2024 ; Zhu *et al*., 2025). Pour les approches basées sur les LLM, les coûts sont mesurés en nombre de jetons et en frais associés (Zhang *et al*., 2025).

Lacunes et limites. Dans l'ensemble, la majorité des travaux existants se concentre sur une seule étape du pipeline d'ER. Plusieurs pipelines sont étroitement liés à des algorithmes ou des implémentations spécifiques, ce qui limite leur généralité. Il manque donc encore un moyen clair et unifié pour représenter, comparer ou assembler des pipelines issus de techniques différentes, ainsi que pour rendre visible la circulation de l'information et la production des décisions à chaque étape. Du point de vue de l'évaluation, les études actuelles restent centrées sur les algorithmes et optimisent des étapes isolées, sans prendre en compte les coûts du pipeline complet. Même lorsque les coûts de calcul sont mesurés, ils sont associés à des opérations individuelles et ne montrent pas comment ces coûts apparaissent et s'accumulent au fil des étapes. Par conséquent, la relation entre performance de l'ER et consommation d'énergie est rarement étudiée, et le compromis performanceénergie des chaînes complètes d'ER demeure largement inexploré, ce qui rend difficile la comparaison de différentes conceptions ou l'évaluation de workflows de bout en bout.

3. Modélisation conceptuelle

Motivés par ces limites, nous passons maintenant de l'identification du problème à l'établissement d'une base conceptuelle pour y remédier. Notre objectif est de rendre le processus ER explicite et analysable, plutôt que de le traiter comme un sous-produit caché d'algorithmes individuels. Pour cela, nous présentons tout d'abord les concepts fondamentaux sur lesquels repose notre cadre.

3.1. Concepts

Nous commençons par formaliser les notions fondamentales qui sont à la base de notre proposition. L'ER est généralement définie comme la tâche consistant à identifier les enregistrements qui font référence à la même entité du monde réel. Formellement :

DÉFINITION 1 (Résolution d'Entité). — *Soit $\mathcal{D} = \{D_1, \ldots, D_m\}$ un ensemble de sources de données, et soit $\mathcal{X} = \bigcup_{j=1}^{m} \mathcal{X}^{(j)}$ l'ensemble de tous les enregistrements. Une relation binaire $R \subseteq \mathcal{X} \times \mathcal{X}$ identifie les enregistrements coréférents, où $(x_i, x_j) \in R$ indique que x_i et x_j se réfèrent à la même entité. Soit $\mathcal{P}$ l'ensemble des processus de calcul opérant sur un domaine de données. La ER est un processus spécifique $\mathcal{P}_{ER} \in \mathcal{P}$ qui calcule cette relation :*

$$\mathcal{P}_{ER} : \mathcal{X} \longrightarrow R,$$

dont l'exécution concrète dépend des paramètres du processus θ spécifiant les choix algorithmiques et stratégiques.

Cette définition précise l'objectif de l'ER, mais n'explique pas comment la relation finale R est obtenue. Les systèmes ER génèrent des preuves intermédiaires à partir d'opérations sur des paires d'enregistrements, et ces preuves constituent la base à partir de laquelle R est dérivée. Afin de rendre cette unité opérationnelle explicite, nous introduisons la notion d'ń état de matching ż, qui est généralement implicite dans les systèmes ER existants en tant que représentation intermédiaire.

DÉFINITION 2 (État de Matching). — *Soit $\mathcal{I} = \{(i,j) \mid i,j \in \mathcal{X}, , i \neq j\}$ l'ensemble des paires d'enregistrements. Un espace d'états de matching $\mathcal{M}$ contient un état m pour chaque $(i,j) \in \mathcal{I}$, initialisé par,*

$$m^{(0)} = ((i,j), t_0),$$

où t_0 est son instant de création. Au cours de l'exécution de l'ER, m accumule des attributs horodatés

$$\mathsf{attr}(m) = ((t_1, v_1), (t_2, v_2), \ldots),$$

où chaque valeur v_k peut représenter, par exemple, un score de similarité ou un état de décision. L'ensemble des états de matching fournit les preuves à partir desquelles R est dérivée :

$$\mathcal{M} \models R.$$

En rendant la procédure d'ER explicite au moyen des *états de matching*, l'ER peut être évaluée non seulement par la justesse de son résultat final, mais aussi par le processus de calcul qui le produit. Toutefois, les évaluations existantes prennent rarement ce processus en compte. Elles se concentrent en général sur la qualité du résultat et estiment le coût du processus à partir du temps d'exécution, qui ne donne qu'une vision partielle et masque la manière dont les preuves s'accumulent au fil des étapes, par exemple à travers les différents degrés de similarité calculés à chaque phase. À l'inverse, la consommation d'énergie est mesurable directement au cours de l'exécution et reflète le coût physique du calcul. Cela permet la notion de *résolution d'entités sensible à l'énergie*.

DÉFINITION 3 (Résolution d'Entités Sensible à l'Énergie). — *Soit $\mathcal{P}_{ER}$ un processus qui calcule une relation R à partir de $\mathcal{X}$ et produit des états de matching $\mathcal{M}$. Sous une configuration θ et un profil matériel $\mathcal{HD}$, l'exécution donne*

$$R = \mathcal{P}_{ER}(\mathcal{X}, \theta), \qquad \mathcal{M} \models R,$$

dont la qualité est évaluée par

$$Perf(R, \mathcal{M}) = \langle Perf_R(R),\ Perf_M(\mathcal{M}) \rangle \in \mathbb{R}^2_{\geq 0},$$

et dont la consommation d'énergie est mesurée par

$$Energy(\mathcal{P}_{ER}, \theta, \mathcal{X}, \mathcal{HD}) \in \mathbb{R}_{\geq 0}.$$

Lorsque l'énergie devient une dimension d'évaluation explicite, l'ER n'est plus optimisée selon un seul objectif. Différentes configurations du processus θ induisent différentes combinaisons de performance et de consommation d'énergie, ce qui fait du choix de θ un problème de décision plutôt qu'un simple choix de conception. Cela conduit naturellement à la notion formelle de *compromis performance-énergie*.

DÉFINITION 4 (Compromis Performance-Énergie). — *Soit $\mathcal{V}$ un espace de valeurs codant les préférences de performance et les tolérances énergétiques, et $\mathcal{C}$ un espace de coûts correspondant aux paramètres configurables influençant $\mathcal{P}_{ER}$ (par exemple, la taille des fenêtres, les fonctions de similarité, la complexité des modèles). Pour une configuration $\theta \in \mathcal{C}$, une fonction d'utilité*

$$U : \mathbb{R}^2_{\geq 0} \times \mathcal{V} \to \mathbb{R}$$

évalue le résultat $(Perf, Energy)$ obtenu. La configuration optimale, sensible à l'énergie, est alors

$$\theta = \arg\max_{\theta \in \mathcal{C}} U\big(Perf(\mathcal{P}_{ER}, \theta, \mathcal{X}), Energy(\mathcal{P}_{ER}, \theta, \mathcal{X}, \mathcal{HD}), \mathcal{V}\big).$$

Ces définitions établissent la base conceptuelle de la résolution d'entités sensible à l'énergie. En nous appuyant sur celles-ci, nous introduisons maintenant un cadre qui organise le processus d'ER en composants modulaires et intègre la prise en compte de l'énergie dans leur exécution.

3.2. Modélisation

En nous appuyant sur ces fondements conceptuels, nous présentons ensuite un cadre unifié qui intègre la résolution d'entités avec une prise en compte explicite de l'énergie. Comme l'illustre la Figure 1, ce cadre s'articule autour de deux parties principales : le processus de résolution d'entités et le mécanisme de sensibilisation à l'énergie. La partie de processus de résolution d'entités présente l'architecture opérationnelle centrale de l'ER, composée de la couche de processus de matching et de la couche de gouvernance du matching. Le mécanisme de sensibilisation à l'énergie intègre la consommation d'énergie comme l'un de ses critères d'évaluation et cherche à trouver un équilibre entre performance et consommation énergétique. Ce mécanisme se compose de la mesure d'énergie et du reporting d'énergie.

Comme le workflow de la couche de processus de matching est déjà résumé à la Section 2 (prétraitement, génération de candidats, comparaison par pairs et comparaison collective), les paragraphes suivants présentent principalement la couche de gouvernance du matching et le mécanisme de sensibilisation à l'énergie.

3.2.1. Couche de gouvernance du matching

Toutes les sorties produites à ces étapes, y compris les paires candidates, ainsi que les scores ou décisions issus du matching pair à pair et collectif, sont progressivement

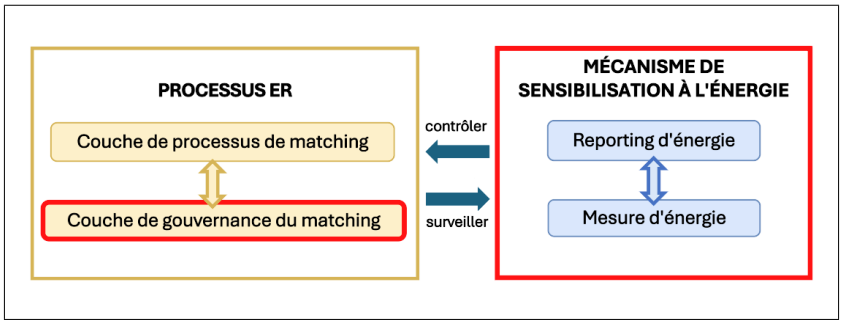


FIGURE 1. *Vue d'ensemble du cadres*

enregistrées dans l'état de matching. Pour gérer le flux d'information et les états intermédiaires entre les étapes, nous introduisons une couche de gouvernance du matching. Son rôle principal est de contrôler, mettre à jour et valider l'état de matching tout au long du pipeline. Elle comprend deux modules : Gestion des États de Matching, qui maintient une vue cohérente et temporelle de tous les résultats, et Évaluation et Rapport de Performance, qui mesure la qualité de l'ER.

Le module de *Gestion des États de Matching* pilote l'état global produit par la couche de processus de matching. Afin d'éviter les incohérences pouvant apparaître lors du traitement en plusieurs étapes, nous intégrons un module d'Intégration des Résultats de Matching (Matching Results Integration, MRI). Un module de Validation et de Restauration d'État (State Commit and Recovery, SCR) est également introduit pour organiser l'historique et assurer des transitions d'état stables. La notion d'*état de matching* est implémentée dans ce module. Chaque instance est créée lors de la Génération de candidats et peut être mise à jour lors du matching par Pairs et du matching Collectif.

Globalement, le composant de Gestion des États de Matching partitionne l'espace $\mathcal{M}$ en sous-ensembles opérationnels correspondant aux différentes phases du cycle de vie d'un état de matching. En particulier, chaque état peut être soit validé, soit encore en cours de traitement. Nous notons ces sous-ensembles comme suit :

$$\mathcal{M}^s \subseteq \mathcal{M} \quad \text{états engagés et stables,}$$
$$\mathcal{M}^a \subseteq \mathcal{M} \quad \text{états actifs maintenus pendant le traitement.}$$

À l'intérieur de $\mathcal{M}^a$, le sous-ensemble des états nouvellement créés ou mis à jour est noté $\mathcal{M}^{\Pi}$; il représente les changements incrémentaux qui déclenchent chaque nouvelle étape de traitement. L'intersection $\widehat{\mathcal{M}}^a = \mathcal{M}^a \cap \mathcal{M}^s$ désigne les états à la fois actifs et déjà validés, constituant l'ensemble de travail stable utilisé lors de l'intégration. Ensemble, les ensembles $\mathcal{M}^a$, $\mathcal{M}^{\Pi}$ et $\mathcal{M}^s$ décrivent tout le cycle de vie d'un état de matching, depuis sa création, en passant par les raffinements successifs, jusqu'à sa stabilisation.

Pour représenter l'arrivée de nouvelles données ou de nouvelles inférences, nous utilisons les symboles t et $t+1$ pour désigner deux étapes successives, où t précède $t+1$. Dans un traitement continu, cela correspond à une progression temporelle ; dans un cadre discret ou par lots, ces symboles représentent simplement l'ordre des étapes.

Le composant *MRI* fait évoluer l'état de matching $\mathcal{M}$ en intégrant les nouvelles preuves produites. Étant donnés l'ensemble actif et validé $\widehat{\mathcal{M}^a}_t$ et l'incrément $\mathcal{M}_t^{\Pi}$ généré par les modules amont, ce composant produit un instantané provisoire de l'état suivant :

$$MRI : (\widehat{\mathcal{M}^a}_t, \mathcal{M}^{\Pi}) \longrightarrow \tilde{\mathcal{M}}^s{}_{t+1},$$

qui respecte les contraintes structurelles et intègre de manière unifiée les mises à jour.

Le composant *SCR* valide cet instantané provisoire et produit un état stable :

$$SCR : \tilde{\mathcal{M}}^s{}_{t+1}, \longrightarrow \mathcal{M}^s_{t+1},$$

SCR assure des validations atomiques et des mécanismes de retour arrière, garantissant qu'aucune mise à jour ne corrompe l'historique, que les échecs partiels ne propagent pas d'erreurs et que les traitements incrémentaux restent cohérents avec les exécutions par lots.

Ensemble, ces deux composants définissent le cycle de vie des états de matching :

$$\mathcal{M}^s_{t+1} = SCR\big(MRI(\mathcal{M}^s_t, \mathcal{M}^a_t, \mathcal{M}^{\Pi}_t)\big).$$

MRI applique les contraintes structurelles et consolide les nouveaux résultats, tandis que SCR garantit des mises à jour atomiques et tolérantes aux pannes. Fondées sur l'état de matching, ces opérations assurent un contrôle de version explicite en mode batch et une cohérence temporelle en mode incrémental, maintenant un état global stable tout au long du pipeline d'ER.

Le module *Évaluation et Rapport de Performance* complète la couche de gouvernance du matching en fournissant un mécanisme standard pour mesurer la qualité du processus d'ER. Les travaux antérieurs utilisent des méthodes d'évaluation différentes selon les approches d'ER. Nous intégrons et généralisons ces stratégies afin que le système puisse adapter automatiquement les procédures d'évaluation les plus appropriées aux méthodes mises en œuvre dans le pipeline.

3.2.2. Mécanisme de sensibilisation à l'énergie

Dans cette partie, nous formalisons le mécanisme de sensibilisation à l'énergie en deux modules : la Mesure d'Énergie (ME) et le Reporting d'Énergie (RE). La ME quantifie l'énergie consommée par un processus d'ER sous une configuration donnée, introduisant ainsi l'énergie comme une dimension d'évaluation explicite et comparable. Le RE réinjecte ce coût énergétique quantifié dans l'espace des configurations, permettant d'ajuster l'exécution en tenant compte de l'énergie. Ensemble, ces modules garantissent que l'énergie est observée pendant l'exécution et qu'elle participe au compromis performance-énergie.

Mesure d'Énergie (ME) Étant donnée une configuration d'ER $\theta \in \mathcal{C}$ et un environnement matériel $\mathcal{HD}$, l'énergie consommée lors de l'exécution de $\mathcal{P}_{ER}$ est obtenue par un module de *ME* :

$$ME : (\mathcal{P}_{ER}, \theta, \mathcal{HD}) \rightarrow \mathbb{R}_{\geq 0}.$$

Nous notons la consommation mesurée par

$$E(\theta) = ME(\mathcal{P}_{ER}, \theta, \mathcal{HD}).$$

En produisant $E(\theta)$ comme une valeur scalaire explicite, l'énergie devient une partie intégrante de l'espace de coûts de l'ER, ce qui permet de comparer les configurations de $\mathcal{C}$ non seulement selon leurs performances, mais aussi selon leurs dépenses en ressources.

Reporting d'Énergie (RE) Alors que la ME rend le coût d'exécution observable, le RE permet à ce coût d'influencer les configurations futures de l'ER. Étant donné l'état de matching $\mathcal{M}$ et l'espace de valeurs $\mathcal{V}$, le *RE* est formalisé par :

$$RE : (E(\theta), \mathcal{M}, \mathcal{V}) \rightarrow \theta',$$

où $\theta' \in \mathcal{C}$ représente une configuration mise à jour qui aligne le comportement énergétique observé avec la sémantique du processus et les préférences de l'utilisateur. La règle de mise à jour de la configuration est donc exprimée comme suit :

$$\theta_{t+1} = RE(E(\theta_t), \mathcal{M}, \mathcal{V}).$$

Ici, $\mathcal{M}$ fournit la sémantique nécessaire pour interpréter si un coût énergétique observé résulte d'un progrès réel du matching ou d'un calcul évitable. Par ce mécanisme, la consommation d'énergie est transformée en information exploitable qui guide l'évolution du système dans l'espace des coûts, lui permettant de s'approcher de la frontière du compromis performanceénergie plutôt que de se limiter à un ajustement uniquement fondé sur la performance.

Dans notre formulation, la configuration θ couvre des paramètres de plusieurs étapes du pipeline d'ER, y compris les règles de prétraitement, les stratégies de génération de candidats, les fonctions de similarité pair à pair et les paramètres d'inférence collective. Chaque étape consomme des réglages dépendant de θ et contribue donc différemment à la performance et à l'énergie, faisant de chaque étape, en principe, un point possible d'adaptation sensible à l'énergie. L'ensemble du pipeline est présenté dans l'Algorithme 1. L'algorithme s'arrête lorsque le système ne reçoit plus de nouvelles données pendant une certaine période, ou sur intervention manuelle de l'utilisateur.

Algorithme 1 : Energy-Aware Entity Resolution

Input : Raw data $\mathcal{X}$, initial configuration $\theta_0 \in \mathcal{C}$, hardware profile $\mathcal{HD}$, preference $\mathcal{V}$

Output : Final entity matching result R, performance report Perf, energy reports $E(\theta_t)$

Initialization;
$t \leftarrow 0$;
$\mathcal{M} \leftarrow \emptyset$;

repeat

1. Activate Energy Measurement;
Start ME for configuration θ_t;

2. Preprocessing;
$\mathcal{X}' \leftarrow \mathrm{Preprocess}(\mathcal{X}, \theta_t)$;

3. Candidate Generation;
$C \leftarrow \mathrm{GenerateCandidates}(\mathcal{X}', \theta_t)$;
$\mathcal{M} \leftarrow \mathrm{MRI.Create}(C)$;

4. Pairwise Matching;
$S \leftarrow \mathrm{ComputePairwiseScores}(C, \theta_t)$;
$\mathcal{M} \leftarrow \mathrm{MRI.Update}(\mathcal{M}, S)$;

5. Collective Matching;
$R \leftarrow \mathrm{CollectiveResolve}(\mathcal{M}, \theta_t)$;
$\mathcal{M} \leftarrow \mathrm{MRI.Update}(\mathcal{M}, R)$;

6. State Commit & Recovery;
$\mathcal{M} \leftarrow \mathrm{SCR}(\mathcal{M}, \theta_t)$;

7. Performance Evaluation;
$\mathrm{Perf} \leftarrow \mathrm{Evaluate}(R, \theta_t)$;

8. Finalize Energy Measurement;
$E(\theta_t) \leftarrow ME(\mathcal{P}_{ER}, \theta_t, \mathcal{HD})$;

9. Energy Reporting and Configuration Adaptation;
$\theta_{t+1} \leftarrow RE\big(E(\theta_t), \mathcal{M}, \mathcal{V}\big)$;
$t \leftarrow t + 1$;

until termination condition is met;
return $(R, \mathrm{Perf}, E(\theta_t))$

4. Implementation

Pour démontrer la faisabilité de notre modélisation conceptuelle, nous mettons en œuvre ses composants principaux dans un prototype modulaire [1], montrant que les pipelines ER existants peuvent être intégrés dans le cadre général et éco-responsable.

Afin de valider la faisabilité et la généralité de notre cadre, nous le mettons en œuvre avec deux modes d'exécution ER distincts : (i) un pipeline d'intégration graphique incrémental, et (ii) un pipeline de matching piloté par un modèle des langues pré-entraîné en mode batch.

4.1. État de matching

Nous implémentons les états de matching comme des objets d'exécution essentiels dans notre prototype. Un état de matching est créé dès qu'une paire d'enregistrements est générée par la chaîne de traitement, puis persiste comme un conteneur explicite d'évidences évolutives, telles que les scores de similarité, les mises à jour de décision et les horodatages. Chaque état progresse à travers différentes phases de son cycle de vie (`pending`, `active`, `committed`) et peut être inspecté, mis à jour ou réutilisé par des étapes d'inférence ultérieures. Une fois validé (`committed`), un état reste disponible comme preuve durable et peut ensuite être archivé sans perdre sa provenance. Cette implémentation montre que les états de matching peuvent être introduits sans modifier la logique existante d'ER. Tout pipeline produisant des paires d'enregistrements ou des scores de similarité peut les alimenter et les faire évoluer. En conséquence, les états de matching ne sont pas des artefacts transitoires, mais de véritables entités d'exécution qui rendent explicites et persistantes les décisions intermédiaires, permettant ainsi des formes de raisonnement au niveau du processus que les pipelines ER traditionnels ne peuvent pas offrir.

4.2. Compromis performance-énergie

Pour mettre en œuvre les deux dimensions d'évaluation introduites dans notre cadre, nous adoptons le F1 score comme métrique de performance et intégrons la mesure de l'énergie dans l'exécution du processus ER. Nous implémentons le module de Mesure d'Énergie à l'aide d'un outil de mesure énergétique au niveau système, de type wattmètres logiciels [2], qui cumule la consommation totale d'énergie (en joules) du CPU, du GPU et de la mémoire vive à des intervalles d'exécution configurables (Álvarez Valera *et al.*, 2024).

1. https://github.com/Mzhongwei/Energy-Aware-Entity-Resolution.git
2. https://github.com/hhumbertoAv/ecofloc.git

Étant donné que la performance et l'énergie opèrent à des échelles différentes, nous normalisons la consommation d'énergie mesurée dans un intervalle unitaire

$$E'(\theta) = \frac{E(\theta)}{\max_{\theta' \in \mathcal{C}} E(\theta')}.$$

Dans un espace de configuration $\mathcal{C}$ qui correspond à l'ensemble des paramètres configurables, pour chaque configuration $\theta' \in \mathcal{C}$, une consommation $E(\theta)$ est mesurée, puis le terme $\max_{\theta' \in \mathcal{C}} E(\theta')$ est déterminé comme la valeur maximale observée sur l'ensemble des configurations évaluées. Cette normalisation exprime ainsi chaque consommation comme une proportion du cas le plus énergivore, garantissant $E'(\theta) \in [0, 1]$ et assurant la comparabilité avec la métrique de performance.

Pour mettre en évidence la tension entre ces deux dimensions, nous exprimons leur compromis à l'aide d'une formulation d'utilité linéaire :

$$U(\theta) = \alpha \cdot Perf_R(R) - \beta \cdot E'(\theta), \qquad \alpha, \beta \in [0, 1],\ \alpha + \beta = 1,$$

où α et β reflètent les préférences de l'utilisateur. Une valeur plus élevée de α met l'accent sur la performance du matching, tandis qu'une valeur plus élevée de β pénalise davantage la dépense énergétique. La valeur d'utilité n'est pas conçue comme un score absolu, mais comme un moyen de classer différentes configurations, permettant d'ordonner les paramètres du processus selon des préférences performanceénergie.

Ainsi, cette fonction d'utilité montre que l'énergie devient une dimension de décision contrôlable et comparable. Les configurations atteignant de meilleures performances peuvent entraîner une consommation énergétique plus élevée, et inversement.

4.3. Transition de paramètres sensible à l'énergie

Dans cette section, nous présentons une instanciation du compromis performance-énergie à travers une adaptation des paramètres, tels que le choix de la méthode de génération de candidats, tout en prenant en compte la consommation d'énergie. Les changements de paramètres sont guidés par des annotations énergétiques plutôt que par des objectifs de performance. Lorsque l'énergie mesurée dépasse un seuil τ défini dans l'espace des valeurs $\mathcal{V}$, le système bascule vers une configuration alternative θ (par exemple, en sélectionnant d'autres fonctions de similarité ou stratégies de matching). Le seuil τ encode des préférences énergétiques définies par l'utilisateur telles qu'un budget énergétique maximal ou une relation acceptable entre performance et énergie et les rend explicites dans la sémantique d'exécution.

Ce mécanisme ne cherche pas à calculer une configuration optimale θ^*. Il démontre plutôt que la résolution d'entités sensible à l'énergie peut exposer l'information énergétique comme une composante du processus d'exécution et prendre en charge des transitions de paramètres sans redémarrer la chaîne de traitement. La fonction d'utilité $U(\theta)$ fournit un moyen formel d'exprimer les préférences, tandis que le seuil détermine uniquement le moment où une transition devient sémantiquement admissible et non le moment où elle est globalement optimale.

Combinées à l'état de matching $\mathcal{M}$, qui stocke les évidences accumulées, les transitions de paramètres préservent les informations déjà calculées. Le pipeline ER poursuit ainsi son exécution sous une autre configuration, produisant des comportements performanceénergie distincts sans réinitialisation. Cela illustre que la résolution d'entités sensible à l'énergie considère les transitions de paramètres joue un rôle important, et ne sont pas simples effets secondaires implicites en dehors du modèle ER.

5. Conclusion

La modélisation que nous proposons établit un cadre conceptuel générique et modulable pour l'ER, en considérant les différentes techniques d'ER comme des instanciations spécifiques d'un même modèle de processus. En rendant explicites les cycles de vie des états de matching, notre solution précise à quels moments les informations doivent être conservées, mises à jour ou éliminées. En liant ce mécanisme à une mesure explicite de l'énergie, l'évaluation de l'ER ne se limite plus à la qualité des résultats obtenus. La consommation énergétique devient alors un paramètre structurant du processus. Nos expérimentations montrent qu'il est possible de représenter conjointement la performance et la consommation d'énergie au sein d'un même cadre d'exécution, et d'exposer explicitement leur compromis à travers la fonction d'utilité proposée. Elles confirment que l'intégration de l'énergie comme dimension de décision peut être prise en compte dans l'exécution des pipelines d'ER, tout en conservant l'historique des inférences et la provenance des décisions dans l'état de matching. Cette représentation conjointe de la performance et de l'énergie ouvre de nouvelles perspectives pour la configuration des pipelines d'ER. Cette représentation rend notamment envisageable, à terme, l'automatisation de cette configuration au sein d'un workflow inspiré de l'AutoML, dans lequel les paramètres seraient ajustés automatiquement afin d'arbitrer entre précision et consommation énergétique, plutôt que d'être définis manuellement.

Bibliographie

Álvarez Valera H. H., Maurice A., Ravat F., Song J., Roose P., Vallès-Parlangeau N. (2024, avril). Energy Measurement System for Data Lake. In *ACIIDS 2024 - 16th Asian Conference on Intelligent Information and Database Systems*. Ras Al Khaimah, United Arab Emirates.

Arora A., Dell M. (2024). Linktransformer: A unified package for record linkage with transformer language models. In *Annual meeting of the association for computational linguistics*, p. 221231.

Baraldi A., Buono F. D., Guerra F., Paganelli M., Vincini M. (2023). An intrinsically interpretable entity matching system. In *International conference on extending database technology*, p. 645–657.

Bienvenu M., Cima G., Gutiérrez-Basulto V., Ibáñez-García Y. (2023, septembre). Combining global and local merges in logic-based entity resolution. In *International conference on principles of knowledge representation and reasoning*, p. 742746.

Binette O., Steorts R. C. (2022, mars). (almost) all of entity resolution. *Science Advances*, vol. 8, n° 12, p. eabi8021.

Christophides V., Efthymiou V., Palpanas T., Papadakis G., Stefanidis K. (2020, décembre). An overview of end-to-end entity resolution for big data. *ACM Comput. Surv.*, vol. 53, n° 6.

Ding H., Dai C., Wu Y., Ma W., Zhou H. (2024). Setem: Self-ensemble training with pre-trained language models for entity matching. *Knowledge-Based Systems*, vol. 293, p. 111708.

Dou W., Shen D., Zhou X., Nie T., Kou Y., Cui H. *et al.* (2023). Soft target-enhanced matching framework for deep entity matching. In *Proceedings of the thirty-seventh aaai conference on artificial intelligence and thirty-fifth conference on innovative applications of artificial intelligence and thirteenth symposium on educational advances in artificial intelligence.*

Fagin R., Kolaitis P. G., Lembo D., Popa L., Scafoglieri F. (2023, septembre). A framework for combining entity resolution and query answering in knowledge bases. In *International conference on principles of knowledge representation and reasoning*, p. 229239.

Gao C., Zhang X., Li L., Li J., Zhu R., Du K. *et al.* (2023, juillet). Ergm: A multi-stage joint entity and relation extraction with global entity match. *Knowledge-Based Systems*, vol. 271, p. 110550.

Gazzarri L., Herschel M. (2023). Progressive entity resolution over incremental data. In *International conference on extending database technology.*

Hu J., Bewong M., Kwashie S., Zhang Y., Nofong V., Wondoh J. *et al.* (2025, juillet). When gdd meets gnn: A knowledge-driven neural connection for effective entity resolution in property graphs. *Information Systems*, vol. 132, p. 102551.

Jabrane M., Tabbaa H., Hadri A., Hafidi I. (2024). Enhancing entity resolution with a hybrid active machine learning framework: Strategies for optimal learning in sparse datasets. *Information Systems*, vol. 125, p. 102410.

Li Y., Li J., Suhara Y., Doan A., Tan W.-C. (2020, septembre). Deep entity matching with pre-trained language models. *Proc. VLDB Endow.*, vol. 14, n° 1, p. 5060.

Liu F., Shen D. (2025, novembre). Token-fusion: A sparse expert routing method for multi-task data matching. In *Acm international conference on information and knowledge management*, p. 49554959.

Maciejewski J., Nikoletos K., Papadakis G., Velegrakis Y. (2025, février). Progressive entity matching: A design space exploration. *International Conference on Management of Data*, vol. 3, n° 1, p. 125.

Moretti A., Shlomo N. (2023, décembre). Improving probabilistic record linkage using statistical prediction models. *International Statistical Review*, vol. 91, n° 3, p. 368394.

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 *International conference on management of data*, p. 1934.

Nanayakkara C., Christen P. (2022, octobre). Locality sensitive hashing with temporal and spatial constraints for efficient population record linkage. In *Proceedings of the 31st acm international conference on information and knowledge management*, p. 43544358.

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.

Papadakis G., Fisichella M., Schoger F., Mandilaras G., Augsten N., Nejdl W. (2023). Benchmarking filtering techniques for entity resolution. In *2023 ieee 39th international conference on data engineering (icde)*, p. 653-666.

Papadakis G., Mandilaras G., Gagliardelli L., Simonini G., Thanos E., Giannakopoulos G. *et al*. (2020, novembre). Three-dimensional entity resolution with jedai. *Information Systems*, vol. 93, p. 101565.

Peeters R., Steiner A., Bizer C. (2024). *Entity matching using large language models.* Consulté sur https://arxiv.org/abs/2310.11244

Ranbaduge T., Vatsalan D., Ding M. (2024, novembre). Privacy-preserving deep learning based record linkage. *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, n° 11, p. 68396850.

Randall S., Brown A., Ferrante A., Boyd J., Robinson S. (2024). Implementing privacy preserving record linkage: Insights from australian use cases. *International Journal of Medical Informatics*, vol. 191, p. 105582.

Wadhwa S., Krishnan A., Wang R., Wallace B. C., Kong L. (2024, novembre). Learning from natural language explanations for generalizable entity matching. In *Conference on empirical methods in natural language processing*, p. 6114–6129.

Wang R., Zhang Y. (2024, juin). Pre-trained language models for entity blocking: A reproducibility study. In *Conference of the north american chapter of the association for computational linguistics*, p. 8720–8730.

Wang T., Chen X., Lin H., Chen X., Han X., Wang H. *et al.* (2024). *Match, compare, or select? an investigation of large language models for entity matching.* Consulté sur https://arxiv.org/abs/2405.16884

Wang Y., Zhou L., Wang Y., Peng Z., Prakash S. (2024, janvier). Leveraging pretrained language models for enhanced entity matching: A comprehensive study of fine-tuning and prompt learning paradigms. *Int. J. Intell. Syst.*, vol. 2024.

Wu R., Chaba S., Sawlani S., Chu X., Thirumuruganathan S. (2020). Zeroer: Entity resolution using zero labeled examples. In *International conference on management of data*, p. 11491164.

Yao D., Gu Y., Cong G., Jin H., Lv X. (2022). Entity resolution with hierarchical graph attention networks. In *International conference on management of data*, p. 429442.

Zeakis A., Papadakis G., Skoutas D., Koubarakis M. (2023, mai). Pre-trained embeddings for entity resolution: An experimental analysis. *Proc. VLDB Endow.*, vol. 16, n° 9, p. 22252238.

Zhang Z., Groth P., Calixto I., Schelter S. (2025). *A deep dive into cross-dataset entity matching with large and small language models.*

Zhu H., Li Z., Jin J. (2025, novembre). GER-LLM: Efficient and effective geospatial entity resolution with large language model. In *Conference on empirical methods in natural language processing*, p. 23272–23288.

Ziyad S., Christen P., Vidanage A., Nanayakkara C., Schnell R. (2025, novembre). Vulnerability-aware hardening for secure privacy-preserving record linkage. In *Acm international conference on information and knowledge management*, p. 45824591.

Reconnaissance d'entités nommées spécifiques - Segmentation & pseudo-annotation

Pape Ibrahima Thiam[1,5], Yohann Chasseray[1,2], Josiane Mothe[1,3], Mathieu Roche[4], Maguelonne Teisseire [5]

1. IRIT, Univ. de Toulouse, CNRS, Toulouse INP, Toulouse, France
pape-ibrahima.thiam,Yohann.Chasseray,Josiane.Mothe
@irit.fr

2. INUC, Toulouse, France

3. CLLE, Univ. de Toulouse, CNRS, Toulouse,

4. TETIS, CIRAD, France, Mathieu.Roche@cirad.fr

5. TETIS, INRAe, France, Maguelonne.Teisseire@inrae.fr

RESUME. *Cet article est une synthèse du travail de Thiam, P. I., Chasseray, Y., Mothe, J., Roche, M., et Teisseire, M., intitulé "Enhancing domain-specific named entity recognition via segmentation and pseudo-labeled annotation". Ce travail a été publié dans les actes de "International Conference on Tools with Artificial Intelligence 2025 (IEEE)" pages 552 à 559, DOI 10.1109/ICTAI66417.2025.00081. Il présente une étude sur l'extraction d'entités nommées avec différentes stratégies de découpage des documents longs ainsi que l'augmentation de données annotées avec des pseudo-labels issus d'un accord inter-extracteur d'entités nommées.*

MOTS-CLÉS : *Reconnaissance d'entités nommées, Environnement à faibles ressources, Segmentation, Traitement Automatique du Langage, Systèmes Alimentaires Territoriaux*

1. Contexte et objectifs

La reconnaissance d'entités nommées (NER) est fondamentale pour structurer l'information textuelle et une étape clé dans de nombreuses tâches de traitement automatique des langues. Les modèles de NER à l'état de l'art, à schéma ouvert,

comme GLiNER (Zaratiana *et al.*, 2024) et NuNER (Bogdanov *et al.*, 2024), permettent d'extraire des entités indépendamment de leur type, mais la taille limitée de leur fenêtre de contexte ne leur permet pas de traiter de façon satisfaisante les documents longs. Ils restent également moins performants sur des domaines avec peu de données d'entraînement.

Les systèmes alimentaires territorialisés (Duarte *et al.*, 2025), auxquels nous nous intéressons dans le cadre d'une collaboration avec l'UMR Innovation[1] financée par le défi O3T[2], sont un exemple typique de domaines sur lesquels les méthodes d'extraction actuelles échouent de par le vocabulaire spécialisé, les types d'entités spécifiques et la longueur des articles de presse traités.

Dans cet article, nous étudions l'impact de différentes stratégies de segmentation de documents et proposons un pipeline semi-supervisé fondé sur l'accord entre des modèles d'extraction pour améliorer la reconnaissance d'entités nommées dans le cadre de domaines spécifiques.

2. Méthogologie

D'un point de vue méthodologique, nous procédons d'abord à une étape de segmentation des documents permettant de nourrir les modèles par des unités de texte porteuses d'informations et respectant la fenêtre contextuelle des modèles. Quatre méthodes de segmentation sont utilisées. *Chunking* (Devlin *et al.*, 2019) consiste à découper régulièrement le texte concerné avec une longueur de segment définie en amont. *Sliding Window* (Beltagy *et al.*, 2020) ajoute une fenêtre glissante entre les segments afin de s'assurer de la capture du contexte à gauche et à droite d'entités possiblement positionnées aux frontières. *Thematic* (Hearst, 1993) s'appuie sur un encoder bidirectionnel pour extraire les segments affichant la cohérence thématique la plus forte. *Thematic + Retriever* étend la stratégie précédente en sélectionnant parmi les segments définis, ceux qui présentent une proximité sémantique forte avec les types d'entités à identifier.

Pour entraîner les modèles de NER à identifier les entités nommées, nous utilisons un jeu de données annotées manuellement, augmenté de données pseudo-annotées par une stratégie d'inter-accord entre modèles de NER. Les modèles (ici GliNER et NuNER) servent donc à la fois d'annotateurs et d'extracteurs d'entités nommées. Pour la pseudo-annotation, seules les entités pour lesquelles les modèles NER sont d'accord sont retenues. Ces données annotées seront utilisées pour réaliser des réglages fins (*fine-tuning*) des modèles de NER. Un autre jeu de données est également utilisé pour le test. Il a été annoté manuellement en grande partie par des experts du domaine et se compose de 237 articles en français spécifiques au domaine, totalisant 3 696 phrases, 83 058 tokens et 6 569 entités nommées annotées.

1. https://umr-innovation.cirad.fr/recherche/collectifs-de-recherche/dam
2. https://o3t.univ-toulouse.fr/

3. Résultats

L'impact de la segmentation est évalué en précision et rappel pour la détection d'entités nommées avec des stratégies zéro et few-shot. Nous utilisons ces modèles – agnostiques des types d'entités à identifier – car ils permettent de réaliser l'identification des entités sans étiquettes *prédéfinies* en déclarant les types d'entités à l'inférence, ce qui est pertinent pour des domaines métiers spécifiques. Le tableau 1 présente les résultats de l'extraction avec GliNER et NuNER pour les différentes méthodes de segmentation étudiées. L'évaluation zero-shot montre que la segmentation par fenêtre glissante offre une meilleure précision que les autres stratégies sur les données des systèmes alimentaires territoriaux, grâce au chevauchement qui préserve la continuité contextuelle, mais au détriment du rappel. Dans ce domaine spécialisé, la précision est privilégiée. La stratégie de fenêtre glissante est donc utilisée ici pour générer des données annotées basées sur l'accord inter-modèles. Cela permet de filtrer les faux positifs et améliorer les performances (précision de 0.714 à 0.753 pour GliNER et de 0.712 à 0.773 pour NuNER, avec même un léger gain en rappel).

Segmentation Strategy	**GLiNER**			**NuNER**		
	P	**R**	**F1**	**P**	**R**	**F1**
Thematic	0.480	**0.442**	**0.460**	0.483	**0.327**	**0.390**
Thematic + Retriever	0.490	0.424	0.455	0.496	0.316	0.386
Sliding Window	**0.549**	0.388	0.454	**0.563**	0.256	0.352
Chunking	0.426	0.266	0.327	0.453	0.185	0.263

TABLEAU 1. *Précision, Rappel, F1 de GLiNER et NuNER en zéro-shot pour différentes stratégies de segmentation sur 237 articles annotés manuellement.*

4. Conclusion

Ce travail montre que, dans un domaine spécialisé et avec peu de données annotées, la combinaison d'une segmentation de texte (pour traiter les longs documents) et d'un processus semi-supervisé basé sur l'accord de deux modèles de NER permet d'améliorer les performances de la détection d'entités nommées. On remarque également que l'accord GLiNER–NuNER fournit des pseudo-labels de meilleure qualité que ceux produits par un seul modèle. Ce travail s'inscrit dans le cadre du projet AI4AGRI, programme de recherche et d'innovation Horizon Europe de l'Union européenne - convention n° 101079136 et a reçu une financement par le défi O3T "Observation de la Terre et Territoires en Transition", défis clés, Région Occitanie.

Bibliographie

Beltagy I., Peters M. E., Cohan A., « Longformer : The long-document transformer », *arXiv preprint arXiv :2004.05150*, 2020.

Bogdanov S., Constantin A., Bernard T., Crabbé B., Bernard E. P., « NuNER : Entity Recognition Encoder Pre-training via LLM-Annotated Data », *EMNLP*, ACL, p. 11829-11841, 2024.

Devlin J., Chang M.-W., Lee K., Toutanova K., « BERT : Pre-training of deep bidirectional transformers for language understanding », *NAACL*, p. 4171-4186, 2019.

Duarte L. A., Méndez M. R., Muñoz-Rojas J., « Territorial embeddedness of sustainable agri-food systems : A systematic review », *Agroecology and Sustainable Food Systems*, vol. 49, nº 6, p. 948-988, 2025.

Hearst M. A., TextTiling : A quantitative approach to discourse segmentation, Technical report, Citeseer, 1993.

Zaratiana U., Tomeh N., Holat P., Charnois T., « GLiNER : Generalist Model for Named Entity Recognition using Bidirectional Transformer », *in* K. Duh, H. Gomez, S. Bethard (eds), *NAACL*, ACL, Mexico City, Mexico, p. 5364-5376, 2024.

Capitaliser les savoir-faire situés par transcription de pratiques : l'expérimentation de Cursus

Guillaume DECHAMBENOIT [1]

1. Bureau de recherches géologiques et minières
3 Av. Claude Guillemin, 45100 Orléans
45000 Orléans
g.dechambenoit@brgm.fr

RESUME. *Cet article propose une démarche de capitalisation des savoir-faire situés par la transcription de pratiques en processus descriptifs intégrés à un système d'information. Notre contribution méthodologique s'appuie sur un système de notation dédié, centré sur les chaînes d'actions observables et les artefacts médiateurs, conçu pour être directement appropriable par les experts métiers. Ce système est opérationnalisé dans Cursus, un studio web de modélisation (éditeur visuel, versioning, interopérabilité BPMN/XPDL/Mermaid, export PDF/image) assorti d'une assistance optionnelle à la modélisation. L'expérimentation, menée au BRGM en alpha interne dans le cadre de la Carte Géologique, vise la constitution d'un référentiel vivant favorisant transmission, traçabilité et reproductibilité, avec une extension envisagée à d'autres domaines des géosciences d'ici 2027.*

MOTS-CLÉS : *Connaissances tacites • Sémiotique des transactions coopératives • Capitalisation des savoir-faire situés • Modélisation de processus*

1. Introduction

La performance des organisations dépend de plus en plus de la capacité à mobiliser des **savoir-faire situés** et des **connaissances tacites** : des savoirs incarnés dans l'expérience, les routines et les ajustements contextuels, que l'on mobilise souvent sans pouvoir les expliciter complètement (Polanyi, 1966)

Cette dimension devient critique dès lors qu'il s'agit de **transmettre**, **réutiliser** ou **outiller** ces savoirs dans un système d'information, notamment via des démarches d'externalisation et de transformation de connaissances (Nonaka and Takeuchi, 1995). Or, une part importante de la connaissance reste enfermée dans des échanges oraux et des artefacts fragmentaires (notes, croquis, messages), pourtant essentiels à la coordination et à la continuité des pratiques (Zacklad, 2007). L'enjeu n'est donc pas seulement de stocker des contenus, mais de rendre ces traces **actionnables** dans le temps.

Nous adoptons une perspective où l'organisation est un **écosystème d'artefacts** (éphémères ou persistants) inscrit dans des unités de stockage qui conditionnent la traçabilité et la réutilisation (Hutchins, 1995) (Norman, 1991) (Star and Ruhleder, 1996). Les formes de transcription traditionnelles demeurent toutefois peu réutilisables et coûteuses cognitivement (Sweller, 1988).

Notre hypothèse est que le tacite devient plus accessible lorsqu'il se stabilise sous forme de **pratiques observables** (Schatzki, 2001) (Orlikowski, 2000). En transcrivant ce que font effectivement les experts (chaînes d'actions, dépendances, décisions, artefacts), on peut produire des **processus descriptifs** compatibles avec l'ingénierie des processus.

Dans ce cadre, l'article traite de **l'externalisation** des connaissances tacites par transcription de pratiques en processus, puis de leur intégration au cœur d'un applicatif dédié, **Cursus**, afin de constituer **un référentiel vivant** supportant transmission, traçabilité, reproductibilité et évolution continue des méthodes.

Plusieurs familles d'outils abordent partiellement cette problématique. Les **modeleurs BPMN** (Signavio, Camunda Modeler) ciblent des processus prescriptifs et exécutables, présupposant une expertise en ingénierie des procédés. Les **outils de diagramme génériques** (draw.io, Miro, Lucidchart) offrent de la flexibilité visuelle mais manquent de sémantique exploitable. Les **plateformes de gestion de connaissances** (wikis, Confluence) restent essentiellement textuelles et peu adaptées à la représentation de flux. Aucune n'est pensée pour une **transcription de pratiques situées** produite par les experts métiers eux-mêmes : c'est le positionnement spécifique visé par Cursus.

2. Objectifs et contributions

Dans ce contexte, ce papier vise à **extraire et capitaliser** une partie des connaissances tacites en transcrivant des pratiques sous forme de **processus descriptifs**, puis en intégrant ces processus au cœur d'un système d'information dédié (Cursus).

Nos contributions sont les suivantes :

– Proposer un cadrage où le tacite est appréhendé comme **un flux de pratiques et d'artefacts** à transcrire et à articuler, plutôt que comme un « contenu » à capturer.

– Outiller cette transcription via un environnement où les pratiques peuvent être décrites, structurées, versionnées et reliées à leurs artefacts (documents, données, messages, preuves).

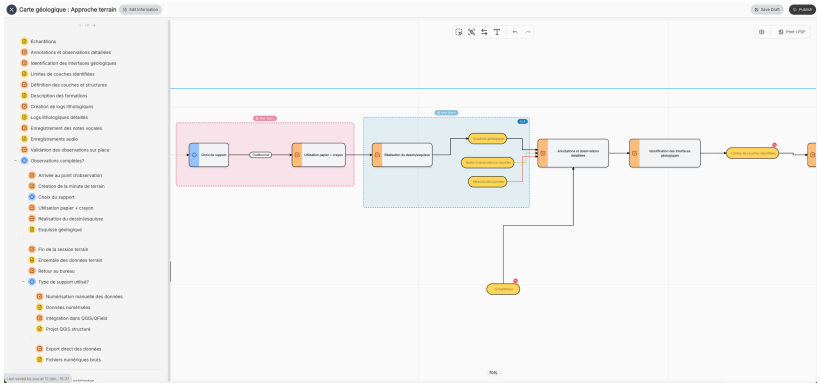


FIGURE 1. *Capture d'écran de l'éditeur de processus de Cursus mettant en avant la modélisation des artefacts médiateurs dans la description d'une pratique d'observation d'un affleurement en géologie.*

– Produire un référentiel vivant de pratiques, destiné à soutenir :

- la transmission et l'apprentissage (onboarding, partage d'expertise),

- la traçabilité et la reproductibilité (notamment dans des contextes scientifiques),

- l'évolution continue des méthodes via la comparaison, la variation et la consolidation des versions.

3. Méthodologie : un système de notation pour transcrire des pratiques

Cursus s'appuie sur un système de notation visuel développé pour transcrire les pratiques expertes (Dechambenoit, 2026).

Plutôt que d'extraire des modèles mentaux embryonnaires, nous ciblons les **pratiques routinisées** (Schatzki, 2001; **?**). La méthode transcrit des **chaînes d'actions observables** plutôt que des pensées (Suchman, 1987; **?**), garantissant un ancrage empirique compatible avec le *process mining*.

Conçu pour être directement appropriable par les experts sans formation en ingénierie (contrairement à BPMN), ce système réduit la charge cognitive (Sweller, 1988) via un langage visuel-spatial simple inspiré de Bertin (Bertin, 1983).

Sa **grammaire minimale** s'articule autour d'une règle stricte : deux tâches ne se connectent jamais directement, mais toujours via un **artefact médiateur**. Elle se compose de six primitives :

– **Contexte/Objectif** (cercle) et **Tâches** paramétrées (rectangles) ;
– **Artefacts** d'information (pilules jaunes) et **Unités de stockage** (pilules noires) ;
– **Espaces** (cadres) et **Flux conditionnels** (décisions/parallélisme).

4. Cursus en pratique : mise en œuvre et expérimentation

4.1. *Mise en œuvre.*

Cursus est un studio web de modélisation construit en **TypeScript** (**Next.js**/React, backend **Firebase**). Il matérialise la grammaire présentée ci-dessus dans un éditeur visuel par glisser-déposer, et porte une définition sémantique des nœuds (tâches, décisions, artefacts, unités de stockage) qui rend les modèles réutilisables. L'outil intègre une bibliothèque de processus avec *versioning*, une assistance contextuelle à la modélisation (*Flowcraft*) et des fonctions d'interopérabilité : import depuis **BPMN** (*Business Process Model and Notation*, OMG 2011), **XPDL** (*XML Process Definition Language*) et **Mermaid** (langage textuel de diagrammes), exports en PDF et image.

4.2. *Expérimentation : production de la Carte Géologique du BRGM.*

Cursus est déployé en alpha interne sur un cas fortement tacite : la **production de la Carte Géologique**, activité collective mêlant levés de terrain, interprétations, harmonisation, validation scientifique et édition. La collecte articule entretiens de praticiens et modélisation directe dans Cursus : chaque pratique est décomposée en chaînes de tâches, rattachée aux artefacts effectivement manipulés (carnets de terrain, photographies, données SIG, cartes de compilation, rapports de validation) et associée à leurs unités de stockage (serveurs métiers, GED, espaces projet). Les premiers processus modélisés alimentent un référentiel vivant, enrichi et versionné par les praticiens eux-mêmes.

4.3. *Accès et perspectives.*

Une **démonstration publique** de Cursus sera proposée en marge de la conférence afin d'en permettre l'évaluation directe ; l'ouverture du code source dans le respect des principes de la Science Ouverte est à l'étude pour les versions ultérieures. À l'horizon **2027**, la méthode et l'outil ont vocation à être étendus à d'autres domaines des géosciences (hydrogéologie, risques, ressources), en conservant le même principe : transcrire des pratiques situées sous forme de processus sémantiquement exploitables et réutilisables.

5. Conclusion

Cette démonstration présente Cursus comme un studio de modélisation qui opérationnalise un système de notation minimal dédié à la transcription de pratiques, en le reliant à un référentiel de processus vivant. L'éditeur visuel, la sémantique des artefacts et le *versioning* offrent aux experts métiers un environnement d'externalisation et de capitalisation des savoir-faire situés, dont l'expérimentation au BRGM autour de la Carte Géologique constitue le premier terrain de validation.

6. Remerciements

Ce travail a bénéficié d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre de France 2030 portant la référence ANR-22-EXSS-0005

Bibliographie

Bertin J., *Semiology of Graphics : Diagrams, Networks, Maps*, University of Wisconsin Press, 1983.

Dechambenoit G., « Bridging Tacit Knowledge and Process Models : A minimal Framework for Practice-Based Process Notation », 2026, Soumis au congrès BPM 2026.

Hutchins E., *Cognition in the Wild*, MIT Press, 1995.

Nonaka I., Takeuchi H., *The Knowledge-Creating Company : How Japanese Companies Create the Dynamics of Innovation*, Oxford University Press, 1995.

Norman D. A., « Cognitive artifacts », *in* J. M. Carroll (ed.), *Designing Interaction : Psychology at the Human-Computer Interface*, Cambridge University Press, p. 17-38, 1991.

Orlikowski W. J., « Using technology and constituting structures : A practice lens for studying technology in organizations », *Organization Science*, vol. 11, n° 4, p. 404-428, 2000.

Polanyi M., *The Tacit Dimension*, Routledge, 1966.

Schatzki T. R., « Practice mind-ed orders », *in* K. Knorr-Cetina, T. R. Schatzki, E. von Savigny (eds), *The Practice Turn in Contemporary Theory*, Routledge, p. 42-55, 2001.

Star S. L., Ruhleder K., « Steps toward an ecology of infrastructure : Design and access for large information spaces », *Information Systems Research*, vol. 7, n° 1, p. 111-134, 1996.

Suchman L. A., *Plans and Situated Actions : The Problem of Human-Machine Communication*, Cambridge University Press, 1987.

Sweller J., « Cognitive load during problem solving : Effects on learning », *Cognitive Science*, vol. 12, n° 2, p. 257-285, 1988.

Zacklad M., « Processus de documentarisation dans les Documents pour l'Action (DopA) », *Hermès, La Revue*, vol. 47, p. 93-110, 2007.

RiskTailor : Automatisation de la personnalisation de rapports de risques cybersécurité par profil utilisateur

Mohamed Abi[1], Abiola Paterne Chokki[2], Jean-François Daune[2]

1. Université Paris-Saclay
9 Rue Joliot Curie, 91190 Gif-sur-Yvette, France
mohamedabi1485@gmail.com

2. Centre d'Excellence en Technologies de l'Information et de la Communication
Avenue Jean Mermoz 28, 6041 Charleroi, Belgique
abiola.chokki@cetic.be, jean-francois.daune@cetic.be

RESUME. *La gestion des risques en cybersécurité constitue un enjeu critique pour les organisations. Les rapports issus des méthodologies standards (ITSRM2, EBIOS RM, ISO 27005) sont souvent d'une complexité qui en limite l'exploitabilité selon le profil des parties prenantes : architectes de sécurité, managers ou experts techniques. RiskTailor est un prototype générant automatiquement des rapports contextualisés et adaptés à chacun de ces profils, en transformant des données structurées issues d'analyses de risques en synthèses accessibles et actionnables grâce au traitement automatique du langage naturel. Une démonstration sur deux cas d'usage industriels valide la pertinence de cette approche.*

ABSTRACT. *Cybersecurity risk management is a critical challenge for organizations. Reports from standard methodologies (ITSRM2, EBIOS RM, ISO 27005) are often too complex to be effectively exploited across stakeholder profiles: security architects, managers, or technical experts. RiskTailor automatically generates contextualized reports tailored to each profile by transforming structured risk analysis data into accessible and actionable summaries using natural language processing. An interactive demonstration on two industrial use cases validates this profile-driven approach.*

MOTS-CLÉS : *cybersécurité, analyse de risques, rapports personnalisés*

KEYWORDS: *cybersecurity, risk analysis, personalized reporting, natural*

1. Motivation

Les rapports d'analyse de risques en cybersécurité constituent un artefact central de la gouvernance des systèmes d'information, comme l'illustrent les cadres méthodologiques EBIOS RM, NIST CSF, ISO 27005 et ITSRM2 (European Commission, 2020). Toutefois, ces méthodologies reposent généralement sur un rapport unique et monolithique destiné à des profils d'acteurs très différents : architectes logiciels, responsables métiers et experts sécurité. Ces acteurs présentent des niveaux de compréhension technique distincts, des priorités décisionnelles différentes et des vocabulaires professionnels spécialisés.

Dans les organisations, l'adaptation manuelle du rapport générique à chaque audience est réalisée ad hoc, par des ajustements et copies de sections. Cette pratique engendre trois difficultés : la densité informationnelle des rapports ralentit l'identification des éléments essentiels, l'absence de contextualisation réduit l'appropriation des recommandations, et cette friction informationnelle allonge les cycles de prise de décision.

Les récents progrès en intelligence artificielle, notamment en traitement automatique du langage naturel et en génération de contenus multi-perspectives, ouvrent la possibilité d'une adaptation systématique et automatisée des livrables de gestion des risques. RiskTailor s'inscrit dans cette perspective : il automatise la transformation de données structurées d'analyse de risques en rapports contextualisés, clairs et optimisés pour différents profils d'utilisateurs, sans modifier les méthodologies de référence ni les données sous-jacentes.

2. Description de RiskTailor

RiskTailor s'articule autour d'une architecture modulaire composée de trois niveaux : entrées, moteur de transformation, et sorties (rapport DOCX).

Entrées. Le système accepte deux catégories de données : un ensemble structuré de données d'analyse de risques au format JSON (produit/système, architecture, composants évalués, risques identifiés, évaluation quantifiée, contre-mesures proposées) ; et la spécification du profil cible et ses préférences. L'utilisateur peut également fournir un template DOCX personnalisé pour la mise en forme.

Moteur de transformation. Le moteur repose sur quatre fonctions modulaires. Le *remplacement dynamique* identifie les variables du template DOCX et les remplace par les éléments correspondants du JSON structuré. La *génération de graphiques* crée des visualisations synthétisant l'architecture du système ou la distribution des risques (diagrammes en secteurs et en barres) adaptées au contexte du profil. La *synthèse textuelle contextualisée* génère des résumés de risques complexes en langage naturel, avec un niveau d'abstraction et un vocabulaire ajustés au profil (technique pour l'expert en sécurité, exécutif pour le manager). Enfin, la *synthèse de listes* produit des listes de

recommandations et de risques filtrées et ordonnées selon la pertinence pour le profil cible.

Les deux dernières fonctions reposent sur le modèle Qwen3 (8B), déployé localement via Ollama, et exploitent un mécanisme de prompts différenciés selon le profil utilisateur afin d'optimiser la qualité de la génération.

Profils utilisateurs. Trois profils utilisateurs ont été définis, chacun associé à un prompt dédié adapté à ses besoins spécifiques. Le profil *Architecte Logiciel* privilégie une approche technique approfondie, mettant en avant les composants affectés, l'impact architectural et les contre-mesures appropriées, avec un vocabulaire spécialisé et un contenu dense centré sur les risques fonctionnels et architecturaux. Le profil *Manager/Directeur Métier* adopte quant à lui un registre exécutif synthétique, orienté vers l'impact business, les coûts, les délais et les risques opérationnels critiques. Enfin, le profil *Expert en Cybersécurité* requiert un contenu exhaustif couvrant les vecteurs d'attaque, les contre-mesures et l'ensemble des détails techniques propres au domaine de la sécurité.

Interface. Le prototype génère un rapport DOCX directement exploitable et s'appuie sur une interface Streamlit (voir Figure 1) pour offrir une interaction intuitive. Une API REST est par ailleurs prévue afin de permettre l'intégration future du système avec les outils GRC (*Governance, Risk & Compliance*).

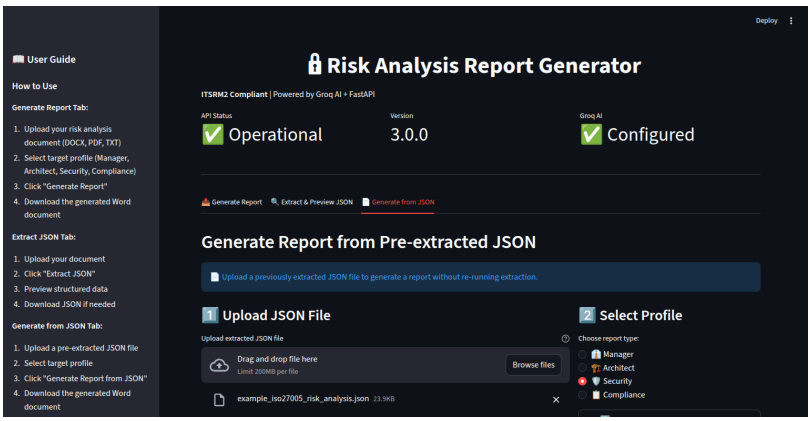


FIGURE 1. *Interface du prototype RiskTailor.*

3. Démonstration interactive

RiskTailor est validé sur deux cas industriels concrets : le robot de surveillance *Securover* et le système de tri *CubiSorter*. Pour chacun, des analyses de risques exhaustives et formalisées ont été menées selon la méthodologie ITSRM2, constituant un ensemble de données d'entrée structuré, cohérent et représentatif.

Flux de démonstration. Chaque participant suit le scénario suivant :

1) *Sélection et chargement* : L'utilisateur sélectionne un cas d'étude (Securover ou CubiSorter) et charge le fichier JSON contenant l'analyse de risques complète.

2) *Configuration du template et du profil* : Choix du template (défaut ou personnalisé) et sélection du profil cible parmi les trois spécialisés.

3) *Génération et comparaison* : Le moteur exécute la transformation en temps quasi–réel. L'utilisateur visualise le rapport généré DOCX.

4) *Validation* : L'utilisateur valide l'adéquation du contenu au profil, la clarté des synthèses et l'absence de bruit informationnel.

Valeur démontrée : La même analyse de risques produit trois rapports radicalement différents. Rapport *manager* : 2-3 pages synthétiques, focus impact business. Rapport *architecte* : 5-6 pages techniques, focus composants. Rapport *expert sécu* : 8-10 pages détaillées, focus vecteurs d'attaque et contre-mesures. Cette comparaison illustre comment la personnalisation améliore l'exploitation de l'information de sécurité selon le rôle du lecteur.

4. Conclusion et perspectives

Les plateformes GRC telles qu'Archer (Archer, 2026) ou LogicGate (LogicGate, 2026) offrent des capacités avancées de structuration des données et de reporting par rôle, mais demeurent centrées sur l'agrégation d'indicateurs sans adapter la narration du risque au profil du destinataire. Les méthodologies EBIOS RM et ITSRM2, bien que rigoureuses dans l'évaluation des risques, n'adressent pas davantage cette dimension communicationnelle. RiskTailor se distingue en automatisant la personnalisation de la narration du risque, améliorant ainsi l'accessibilité de la gouvernance et l'adhésion aux mesures de sécurité recommandées.

Les travaux futurs porteront sur une validation empirique sur des cas d'étude diversifiés, la comparaison de modèles LLM alternatifs, l'intégration de mécanismes d'interaction humaine pour un ajustement itératif des rapports, ainsi que l'automatisation de l'extraction de données à partir de formats PDF et DOCX existants.

Remerciements. Ce travail a été financé par le projet de recherche Wallon CyberExcellence (2110186).

Bibliographie

Archer, https://www.archerirm.com, 2026. Accessed : 2026.

European Commission, « IT Security Risk Management Methodology (ITSRM) », 2020. v1.2.

LogicGate, https://www.logicgate.com, 2026. Accessed : 2026.

RiskMate : Assistant collaboratif humain–IA pour l'atténuation des risques de cybersécurité

Kenza Khemar[1], Abiola Paterne Chokki[2], Thierry Noundou Njike[2], Jean-François Daune[2]

1. Université de Mons
Place du Parc 20, 7000 Mons, Belgique
kenza.khemar@student.umons.ac.be

2. Centre d'Excellence en Technologies de l'Information et de la Communication
Avenue Jean Mermoz 28, 6041 Charleroi, Belgique
abiola.chokki@cetic.be, thierry.noundounjike@cetic.be, jean-francois.daune@cetic.be

RESUME. *L'analyse et l'atténuation des risques cybernétiques reposent sur des jugements d'experts difficiles à standardiser, chronophages et sensibles à l'expérience des analystes. Nous présentons RiskMate, un assistant collaboratif humain–IA destiné à soutenir la prise de décision en gestion des risques de cybersécurité. L'outil structure l'identification des contre-mesures, explicite les justifications associées et permet un ajustement humain traçable des estimations d'impact et de probabilité. RiskMate s'inscrit dans une démarche de gouvernance des systèmes d'information où l'IA assiste sans automatiser la décision. La démonstration interactive montre comment RiskMate accompagne un ingénieur sécurité dans l'élaboration et la justification d'une stratégie d'atténuation de risque sur des cas industriels réels.*

ABSTRACT. *Cybersecurity risk mitigation relies on expert judgement that is difficult to standardize, time-consuming and dependent on analyst experience. We present RiskMate, a collaborative human–AI decision-support assistant designed to support cybersecurity risk management. The tool structures countermeasure identification, provides explicit justifications, and allows traceable human adjustment of likelihood and impact reduction estimates. RiskMate follows an information systems governance perspective where AI assists but does not automate decision-making. The live demonstration shows how RiskMate supports a security engineer in elaborating and justifying mitigation strategies on real industrial cases.*

MOTS-CLÉS : *cybersécurité, gestion des risques, aide à la décision, explicabilité.*

KEYWORDS: *cybersecurity, risk management, decision support, explainability.*

1. Motivation

La gestion des risques de cybersécurité constitue une activité structurante pour la gouvernance des systèmes d'information. Elle implique l'identification de menaces, la sélection de contre-mesures et la justification des décisions auprès d'acteurs organisationnels (par exemple : auditeurs) (Jøsang, 2025). Malgré l'existence de méthodologies reconnues (ISO 27005, EBIOS RM, ITSRM2 (Commission, 2020)), leur mise en oeuvre reste fortement dépendante de l'expertise individuelle, ce qui entraîne hétérogénéité des décisions, délais d'analyse importants et difficulté de justification formelle.

Par ailleurs, l'introduction récente des systèmes d'intelligence artificielle dans les processus décisionnels organisationnels pose des questions de responsabilité, d'auditabilité et de confiance. Dans ce contexte, l'objectif n'est pas d'automatiser la décision de sécurité mais d'assister l'analyste tout en conservant un contrôle humain explicite.

Nous proposons RiskMate, un assistant interactif destiné à structurer l'analyse d'atténuation des risques.

2. Description de RiskMate

RiskMate est un assistant d'aide à la décision reposant sur le modèle Gemma 4 et intégrant une boucle de validation humaine. Le fonctionnement de l'outil s'articule en deux étapes complémentaires.

Identification des contre-mesures. À partir de la description textuelle d'un risque fournie par l'utilisateur, l'outil s'appuie sur le catalogue de contre-mesures d'ITSRM2 comme base de connaissances structurée, afin de contraindre l'espace de génération et de limiter les risques d'hallucination. Il propose ensuite un ensemble de contre-mesures préventives, détectives, correctives et compensatoires adaptées au contexte du cas d'étude, accompagnées de justifications explicitant les critères ayant motivé chaque suggestion. L'utilisateur est ainsi en mesure d'examiner ces propositions et leurs justifications respectives afin de prendre des décisions éclairées quant à la conservation ou au rejet de chacune d'elles.

Estimation de l'atténuation. Une fois les contre-mesures validées, l'outil estime, pour chacune d'elles, les valeurs du facteur de réduction de probabilité (Likelihood Reduction Factor, LRF) et du facteur de réduction d'impact (Impact Reduction Factor, IRF) sur le niveau de risque résiduel, accompagnées d'une justification textuelle détaillée (Commission, 2020). À l'instar de l'étape précédente, l'utilisateur conserve la possibilité de réajuster manuellement ces valeurs selon son jugement expert.

Interface. RiskMate propose une interface multimodale (voir Figure 1) : une interface chatbot, fournie par n8n, permettant d'exprimer les risques en langage naturel ; une API REST, également fournie par n8n, pour l'intégration dans des outils existants ; ainsi qu'un tableau de synthèse récapitulant les contre-mesures et les risques résiduels.

FIGURE 1. *Interface conversationnelle de RiskMate.*

Evaluation préliminaire. Une évaluation préliminaire, conduite avec deux ingénieurs en sécurité sur les deux cas d'étude présentés à la section démonstration, a montré que l'outil était en mesure de formuler des suggestions de contre-mesures pertinentes, d'estimer des valeurs d'atténuation cohérentes et de produire des justifications satisfaisantes. Il a toutefois été observé que, pour certains risques testés, une description plus précise et contextualisée du risque s'avérait nécessaire afin d'obtenir des suggestions spécifiques au cas d'étude plutôt que des recommandations de portée générale.

Comparaison avec les solutions existantes. Les référentiels de gestion des risques tels qu'ITSRM2 ou ISO 27000 proposent des contre-mesures aux menaces fondées sur l'expertise humaine, selon une approche exclusivement manuelle. Des travaux antérieurs (Tao *et al.*, 2025; Al-E'mari *et al.*, 2025) ont exploré des approches assistées par l'intelligence artificielle, sans toutefois se focaliser sur l'évaluation des facteurs d'atténuation ni sur la validation humaine (Karunamurthy, 2023).

RiskMate se distingue par sa capacité à interpréter la description d'un risque fournie par l'utilisateur afin de proposer un ensemble pertinent de contre-mesures, accompagné de coefficients d'atténuation et de justifications explicatives. Il intègre en outre une étape de validation par l'utilisateur, préalablement à l'application des contre-mesures prescrites.

3. Démonstration interactive

La démonstration s'appuie sur deux cas industriels : un robot de surveillance autonome et un système automatisé de tri industriel. Ces cas disposent d'analyses de

risques préalables selon la méthodologie ITSRM2. Lors de la session, un participant jouera le rôle d'analyste sécurité. Le scénario complet dure environ cinq minutes :

1) saisie d'un scénario de risque en langage naturel,

2) génération et affichage des contre-mesures proposées,

3) modification interactive par le participant,

4) recalcul immédiat du risque résiduel et consultation des justifications.

L'utilisateur observe en temps réel l'impact de ses décisions et la justification associée, illustrant l'intérêt de l'assistance pour la documentation et la traçabilité des analyses.

4. Conclusion et perspectives

RiskMate illustre l'intégration d'un assistant d'IA dans un processus organisationnel de gestion des risques sans délégation de la décision. Le système structure l'analyse, facilite la justification et documente les choix de sécurité, contribuant à la gouvernance des systèmes d'information.

Les travaux futurs s'orienteront vers une validation empirique à plus grande échelle, l'exploitation systématique des retours utilisateurs et l'intégration de bases de connaissances complémentaires, avec pour objectif d'accroître la qualité, la couverture et la pertinence des recommandations générées par l'outil.

Remerciements. Ce travail a été financé par le projet de recherche Wallon CyberExcellence (2110186).

Bibliographie

Al-E'mari S., Sanjalawe Y., Fataftah F., « Autonomous Cyber Underwriting Agents Reinventing Risk Evaluation Through Reinforcement Learning and Multi-Agent Systems », *Advances in Computational Intelligence and Robotics*, 2025.

Commission E., IT Security Risk Management Methodology v1.2, Technical report, Directorate General for Digital Services, 2020.

Jøsang A., « Cyber Risk Management », *Cybersecurity*, Springer, Cham, 2025.

Karunamurthy A., « Human-in-the-Loop Intelligence : Advancing AI-Centric Cybersecurity for the Future », *International Journal of Multidisciplinary Scientific Research and Development*, 2023.

Tao J., Li G., Fei J. *et al.*, « Multi-Agent for Network Security Monitoring and Warning : A Generative AI Solution », *IEEE Network*, 2025.

Chaos4CPS : Outil assisté par un agent IA pour la conception d'expériences d'ingénierie du chaos de systèmes complexes

Abiola Paterne Chokki, Christophe Ponsard, Jean-François Daune

Centre d'Excellence en Technologies de l'Information et de la Communication
Avenue Jean Mermoz 28, 6041 Charleroi, Belgique
{abiola.chokki, christophe.ponsard, jean-francois.daune}@cetic.be

RESUME. *Assurer la robustesse des systèmes cyber-physiques est crucial pour garantir leur résilience face à des perturbations inattendues. Bien que l'ingénierie du chaos fournisse une méthodologie structurée, sa phase de conception manque d'un support systématique couvrant la caractérisation du système, la définition des métriques, la formulation des hypothèses et la planification des perturbations. Pour répondre à ce besoin, nous proposons Chaos4CPS, un outil assisté par intelligence artificielle reposant sur une architecture multi-agents dont les agents spécialisés automatisent la conception des expériences tout en produisant des justifications traçables. Cette démonstration présente comment Chaos4CPS accompagne systématiquement chaque phase de l'ingénierie du chaos.*

ABSTRACT. *Ensuring the robustness of cyber-physical systems is crucial to guarantee their resilience against unexpected perturbations. Although chaos engineering provides a structured methodology, its design phase lacks systematic support covering system characterization, metric definition, hypothesis formulation, and perturbation planning. To address this gap, we propose Chaos4CPS, an AI-assisted tool based on a multi-agent architecture in which specialized agents independently automate experiment design while producing traceable justifications. This demonstration presents how Chaos4CPS systematically supports each phase of the chaos engineering methodology.*

MOTS-CLÉS : *robustesse, systèmes cyber-physiques, ingénierie du chaos, système multi-agents*

KEYWORDS: *robustness, cyber-physical systems, Chaos engineering, multi-agent systems*

1. Motivation

La robustesse des systèmes cyber-physiques est un enjeu critique : ces systèmes, où des composants logiciels interagissent en temps réel avec des processus physiques via capteurs et actionneurs, sont exposés à des défaillances aux conséquences potentiellement graves. Des enquêtes menées auprès d'entreprises, notamment des PMEs, révèlent des faiblesses persistantes dans les méthodes de test et un besoin d'approches structurées pour améliorer la résilience (Shah *et al.*, 2016; Ponsard *et al.*, 2025). L'ingénierie du chaos répond à ce besoin en provoquant de manière volontaire et contrôlée des perturbations sur un système en conditions réelles, afin d'en évaluer la résilience. Initialement introduite pour les systèmes cloud (Netflix, 2025), cette méthodologie suit quatre étapes : la caractérisation de l'état stable, la formulation d'hypothèses de comportement (SI-ALORS-PARCE QUE), l'expérimentation par injection de perturbations, et l'analyse des résultats. Son application aux systèmes cyber-physiques soulève cependant des défis spécifiques liés aux couplages cyber-physiques, aux contraintes temps-réel et aux exigences de sûreté, pour lesquels la phase de conception manque d'un support systématique. Pour répondre à ce manque, nous présentons Chaos4CPS (code accessible à l'adresse `https://github.com/cetic/chaos4cps`), un outil qui permet à la fois de conduire manuellement le processus d'ingénierie du chaos étape par étape, et d'automatiser la conception des expériences via une architecture multi-agents assistée par intelligence artificielle, avec ou sans supervision humaine.

2. Description de Chaos4CPS

Chaos4CPS se compose de deux interfaces complémentaires : une interface de gestion du cycle de vie de l'ingénierie du chaos pour l'encodage manuel des données, et une interface conversationnelle implémentant une architecture multi-agents pilotée par l'intelligence artificielle pour automatiser la conception des expériences. L'outil s'appuie sur Django, PostgreSQL et n8n, avec GPT-4o mini comme modèle de langage sous-jacent (température = 0,4). La Figure 1 présente l'interface centralisée.

Architecture multi-agents. Quatre agents d'analyse mettent en oeuvre les phases de l'ingénerie du chaos. L'*agent d'identification des composants* cible les couplages cyber-physiques, les contraintes temps-réel et les éléments critiques pour la sûreté. L'*agent de métriques en état stable* définit des indicateurs de déviation à trois niveaux (Avertissement, Dégradation, Défaillance). L'*agent d'identification des hypothèses* structure les sorties selon un schéma SI-ALORS-PARCE QUE (Figure 2). L'*agent de conception d'expériences* assure une couverture complète des hypothèses et l'intégration des mesures de sécurité. Chacun s'appuie sur trois sous-agents : un *sous-agent d'analyse*, un *sous-agent de révision* (actif en mode avec supervision humaine, permettant à l'utilisateur de valider ou rejeter les suggestions) et un *sous-agent de mise en forme*. Le contexte se propage séquentiellement entre les phases. Tous les agents produisent des champs de *Justification* obligatoires rendant le raisonnement auditable. Les sorties sont persistées dans PostgreSQL et consultables depuis l'interface centralisée.

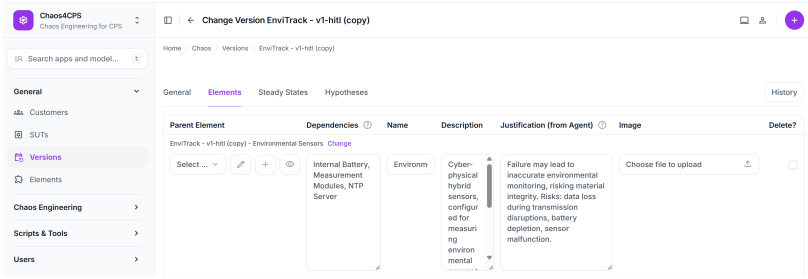


FIGURE 1. *Interface utilisateur de Chaos4CPS.*

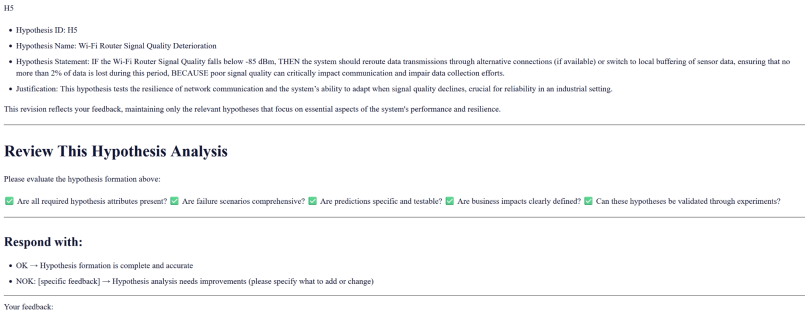


FIGURE 2. *Aperçu de l'agent IA sur la phase d'identification d'hypothèses*

Comparaison avec les travaux connexes. Les techniques traditionnelles de test de robustesse (injection de fautes (Ziade *et al.*, 2014), test par génération aléatoire d'entrées (Manès *et al.*, 2018), test à base de modèles (Moradi *et al.*, 2023)) se concentrent sur l'exécution d'expériences prédéfinies, sans aborder la question en amont de leur conception. Les outils d'ingénierie du chaos existants (Netflix, 2025; Gremlin, 2025; LitmusChaos, 2025) ciblent les systèmes cloud sans adaptation aux systèmes cyber-physiques, tandis que CHAOSEATER (Kikuta *et al.*, 2025) ne propose ni caractérisation du système, ni explicabilité, ni supervision humaine. Chaos4CPS combine de manière originale ces dimensions.

3. Démonstration interactive

La démonstration se déroulera en mode avec supervision humaine. L'utilisateur fournit l'identifiant de son système (préalablement encodé dans Chaos4CPS) ou une description textuelle, puis sélectionne l'agent à activer (1 : Caractérisation du système,

2 : État stable, 3 : Hypothèses, 4 : Conception d'expériences). L'agent génère des suggestions que l'utilisateur affine de manière itérative jusqu'à validation explicite, déclenchant la sauvegarde des données et le passage à la phase suivante. Une fois fini, un rapport complet est généré. Le système EnviTrack est pré-encodé comme cas d'étude de référence pour les participants sans système propre. Il s'agit d'un système de surveillance environnementale pour entreprises agroalimentaires, assurant le suivi des conditions de stockage (température, humidité, gaz) via des capteurs déployés dans des silos, avec une passerelle relayant les données depuis de l'internet des objets vers un système centralisé de surveillance et de génération d'alarmes.

4. Conclusion et perspectives

Chaos4CPS illustre l'apport de l'intelligence artificielle pour la conception systématique d'expériences d'ingénierie du chaos sur les systèmes cyber-physiques, en structurant l'analyse, en justifiant les choix et en documentant les mesures de sécurité. Les travaux futurs porteront sur la génération automatique de scripts d'exécution, le développement de versions spécialisées par domaine et l'intégration de mécanismes d'apprentissage pour affiner les suggestions au fil du temps.

Remerciements. Ce travail a été partiellement financé par les projets de recherche de la Région Wallonne CARAPACE (convention 2310088) et CyberExcellence (convention 2110186).

Bibliographie

Gremlin, https://www.gremlin.com, 2025. Accessed : 2025.

Kikuta D., Ikeuchi H., Tajiri K., « ChaosEater : Fully Automating Chaos Engineering with Large Language Models », *arXiv preprint arXiv :2501.11107*, 2025.

LitmusChaos, https://litmuschaos.io, 2025. Accessed : 2025.

Manès V. *et al.*, « The Art, Science, and Engineering of Fuzzing : A Survey », *IEEE Transactions on Software Engineering*, vol. 47, p. 2312-2331, 2018.

Moradi M., Van Acker B., Denil J., « Failure Identification Using Model-Implemented Fault Injection with Domain Knowledge-Guided Reinforcement Learning », *Sensors*, vol. 23, p. 2166, 2023.

Netflix, https://netflix.github.io/chaosmonkey, 2025. Accessed : 2025.

Ponsard C., Chokki A. P., Daune J.-F., Majchrowski A., « Défis et Besoins des PME pour la Robustesse de leurs Systèmes Cyber-Physiques », 2025. Accessed : 2025.

Shah S. M. A., Sundmark D., Lindström B., Andler S. F., « Robustness Testing of Embedded Software Systems : An Industrial Interview Study », *IEEE Access*, vol. 4, p. 1859-1871, 2016.

Ziade H., Ayoubi R., Velazco R., « A Survey on Fault Injection Techniques », *International Arab Journal of Information Technology (IAJIT)*, vol. 1, n° 2, p. 25-40, 2014.

Adoption des systèmes de recrutement basés sur l'IA : avantages, défis et confiance des utilisateurs

Haoyue Liu[1], Rebecca Deneckère[2]

Centre de Recherche en Informatique, Université Paris 1 Panthéon-Sorbonne
90 rue Tolbiac, Paris, France
1. haoyue.liu@inria.fr
2. rebecca.deneckere@univ-paris1.fr

REFERENCE DE L'ARTICLE RCIS. Ce texte est un résumé de l'article : « Adoption of AI-based Recruitment Systems: Benefits, Challenges, and User Trust », accepté à RCIS 2026.

MOTS-CLES. Intelligence artificielle; recrutement; systèmes d'information; biais algorithmique; perception des utilisateurs; aide à la décision.

1. Introduction

Le processus de recrutement a été profondément transformé par l'adoption croissante de l'intelligence artificielle (Madhavi.K. et al, 2024)(Rajapriya.S. et Subrahmanyam.M., 2024). Autrefois principalement manuel et chronophage, il est désormais soutenu par des outils intelligents capables d'automatiser et d'optimiser plusieurs étapes clés, comme la gestion des candidatures, la présélection, les entretiens et l'aide à la décision (Ben Ayed.H., 2024). Cette évolution s'est accélérée avec la pandémie de COVID-19, qui a renforcé le recours aux pratiques de recrutement numériques et à distance(Rajapriya.S. et Subrahmanyam.M., 2024). Cependant, la littérature reste fragmentée, car peu d'études proposent une vision globale de l'intégration de l'IA dans l'ensemble du processus de recrutement, et les recherches sur la perception des utilisateurs restent limitées(Rigotti, L., Fosch-Villaronga, E, 2024) (Melliani, S., El Kharbouchi, A., 2024). Dans ce contexte, cette étude cherche à répondre à la question suivante : Quelles sont les perceptions des utilisateurs quant à l'utilisation de l'IA tout au long du processus de recrutement ? Pour répondre à cette question, l'étude s'articule autour de quatre sous-questions : (RQ1) Quelles phases du processus de recrutement sont impactées par l'IA ? (RQ2) Quelles techniques d'IA sont utilisées à travers ces phases ? (RQ3) Comment les utilisateurs perçoivent-ils les systèmes de recrutement basés sur l'IA ? (RQ4) Quels sont les défis et les limites associés à l'utilisation de l'IA dans le recrutement ?

2. Résultats

La littérature montre que l'usage de l'IA dans le recrutement varie selon les phases du processus (QR1) (voir Figure1). Elle est surtout utilisée dans les étapes initiales, comme le sourcing et le tri des candidatures, tandis que son rôle reste limité dans la décision finale, où le jugement humain demeure central. Plusieurs techniques sont mobilisées (QR2), notamment le machine learning pour l'analyse et le classement des CV, le traitement automatique du langage naturel pour l'extraction d'informations et les échanges automatisés, ainsi que la vision par ordinateur dans les entretiens vidéo. Globalement, l'IA est mieux acceptée dans les tâches standardisées et volumineuses que dans les étapes plus sensibles. Les perceptions des utilisateurs (QR3), analysées à partir de deux méthodologies complémentaires, un Survey auprès des candidats et des entretiens avec des recruteurs, montrent une acceptation nuancée. Les candidats reconnaissent des gains de rapidité et d'efficacité, mais expriment des inquiétudes sur l'équité, la transparence et la prise en compte des profils atypiques (QR4). Les recruteurs, quant à eux, considèrent surtout l'IA comme un outil d'aide à la décision, utile pour réduire la charge de travail, tout en restant prudents quant à son autonomie dans les entretiens et les décisions finales. Dans l'ensemble, son acceptation dépend de sa transparence, de son contrôlabilité et du maintien d'une supervision humaine.

Rôle IA	Besoins du poste et offre d'emploi	Tri des candidats	Interviews	Sélection finale
Automatiser	Rédaction des offres d'emploi	Analyse des CV	Chatbots pour les questions fréquentes	Synthèse des informations probantes
	Extraction des compétences (NLP)	Extraction des entités et des compétences	Soutien à la planification	Génération de rapports
	Reformulation du contenu (IA générative)	Élimination des doublons	Transcription des entretiens	Documentation de la justification des décisions
Recommander/Classer	Suggestion de mots-clés	Évaluation et classement des candidats	Mise en évidence d'indicateurs	Mise en évidence des risques ou biais potentiels
	Recommandation des canaux de diffusion	Soutien à la présélection	Signalement des incohérences	Comparaison des candidats finalistes
	Références du marché	Mise en correspondance des candidats avec les descriptions de poste	Suggestion de questions de relance	Explication des compromis
Évaluer	Évaluation à faible enjeu	Prédiction de l'adéquation candidat-poste	Analyse des indices vidéo ou vocaux	Légitimité limitée
	Principalement axée sur la qualité textuelle	Inférence de signaux faibles	Extraction de signaux comportementaux	Utilisation avec prudence
		Traitement sensible aux biais	Niveau élevé d'incertitude	Nécessité d'un contexte humain
Aider à la décision	Validation humaine	Supervision humaine	Jugement humain	Décision humaine
		Mécanismes de recours ou de réexamen	Évaluation contextuelle	Responsabilité explicite

Conditions de confiance			
Besoins du poste et offre d'emploi	**Tri des candidats**	**Interviews**	**Sélection finale**
Transparence : faible à moyenne (assistance à la rédaction)	Transparence : élevée (pourquoi le candidat a été retenu ou rejeté)	Transparence : élevée (éléments mesurés)	Responsabilité : responsabilité humaine explicite
Contrôle : élevé (résultats modifiables)	Contrôle : possibilité de correction et de révision du processus	Confidentialité : consentement et minimisation des données	Contestabilité : justification et documentation
Équité : surveiller les biais de formulation	Équité : suivi des biais et audits	Validité : éviter la surinterprétation	Gouvernance : clarté des rôles et traçabilité

Figure 1 : Rôle de l'intelligence artificielle aux différentes étapes du recrutement.

3. Discussion

Cette recherche montre une double perception de l'IA dans le recrutement. Les résultats montrent, d'un côté, une acceptation pragmatique de l'IA par les candidats, principalement liée aux gains d'efficacité, et, de l'autre, une acceptation plus conditionnelle chez les recruteurs, davantage associée aux enjeux de gouvernance. Cette différence souligne que l'acceptation de l'IA dépend du rôle des utilisateurs, de

leurs attentes et du niveau de contrôle qu'ils conservent sur les processus décisionnels. Dans cette optique, les résultats soutiennent l'idée de modèles de recrutement hybrides, dans lesquels l'IA est utilisée comme un outil d'aide à la décision plutôt que comme un décideur autonome.

Sur le plan théorique, cette étude souligne le rôle central de la confiance, de la transparence et de la supervision humaine dans l'acceptation de l'IA en ressources humaines (Rigotti, L. et al, 2024). Sur le plan pratique, les résultats suggèrent que les organisations devraient privilégier l'explicabilité, une communication claire et une gouvernance éthique lors du déploiement de systèmes de recrutement fondés sur l'IA (Thalpage,N.S. et al, 2023). Ce travail met ainsi en lumière une exigence sociotechnique essentielle : aligner les capacités des systèmes avec des mécanismes de gouvernance adaptés afin de préserver la confiance et la responsabilité (Choung, H.et al, 2023).

De futures recherches pourraient prolonger ce travail en étudiant l'évolution de l'adoption de l'IA dans le temps, les différences culturelles, ainsi que l'effet de l'explicabilité sur la confiance et l'acceptation dans le recrutement.

Bibliographie

Madhavi, K., Kaveri, S.: AI Integration in HR Processes. In: Balancing Automation and Human-Machine Interaction in Business, pp. 55–72. IGI Global, Hershey, PA (2024).

Rajapriya, S., Subrahmanyam, M.: A Conceptual Study on Perception Towards the Implementation of Artificial Intelligence in the Recruitment and Selection Process in MNC Companies. Balancing Automation and Human-Machine Interaction in Modern Marketing, pp. 1–18. IGI Global, Hershey, PA (2024).

Rigotti, L., Fosch-Villaronga, E.: Fairness, AI and recruitment. Computer Law & Security Review 52, 105966 (2024).

Ben Ayed, H.: Recruitment in the Era of Web 4.0: What Place for Artificial Intelligence in Recruitment? Qubahan Academic Journal 4(2), 45–58 (2024).

Rigotti, L., Fosch-Villaronga, E.: Artificial Intelligence Employment Interviews: Examining Limitations, Biases, and Perceptions. Computer 57(6), 78–87 (2024). IEEE.

Melliani, S., El Kharbouchi, A.: Impact of E-Recruitment on Recruiter-Candidate Engagement. SA Journal of Human Resource Management 19(3), 1–12 (2024).

Choung, H., David, P., Ross, A.: Trust in AI and Its Role in the Acceptance of AI Technologies. International Journal of Human–Computer Interaction 39(5), 1017–1032 (2023).

Thalpage, N. S.: Unlocking the Black Box: Explainable Artificial Intelligence (XAI) for Trust and Transparency in AI Systems. AI and Ethics 4(1), 33–48 (2023).

Proposition d'un trade-off éco-responsable pour l'évaluation d'algorithmes d'analyse dans les graphes temporels

Pierre-Paul Cavallera[1], Landy Andriamampianina[1], Moncef Garouani[1], Jiefu Song[1], Franck Ravat[1], Nathalie Vallès-Parlangeau[2]

1. IRIT, Université Toulouse Capitole, 2 rue du doyen Gabriel Marty, F-31042 Toulouse Cedex 9, prenom.nom@irit.fr

2. LIUPPA, Université de Pau et des Pays de l'Adour, Avenue de l'Université, F-64000 Pau, prenom.nom@univ-pau.fr

RESUME. *L'analyse prédictive de graphes temporels est cruciale pour comprendre des comportements spatio-temporels complexes, mais son utilité est souvent compromise par des coûts de calcul prohibitifs, qui engendrent des coûts énergétiques élevés. Pour intégrer la dimension d'éco-responsabilité, notre démarche s'appuie sur deux leviers : l'optimisation de la puissance matérielle afin de réduire le coût énergétique et le temps d'exécution des calculs, et un compromis d'évaluation entre efficacité, coût et performance. À l'inverse des méthodes classiques centrées sur la précision, nous adoptons une analyse multi-objectifs qui corrèle efficacité, coût énergétique et performance selon des configurations matérielles données. De plus, nous évaluons le passage à l'échelle en éprouvant le système à des jeux de données de natures spatiales ou temporelles différentes. Nos résultats démontrent qu'une optimisation axée sur la réduction du coût énergétique peut permettre de réduire la consommation énergétique jusqu'à 3,4 fois au prix d'une augmentation marginale de 7 % de l'erreur de prédiction, soulignant ainsi le potentiel de réductions du coût énergétique significatives. En offrant une analyse du compromis entre performance et énergie, ces travaux fournissent aux chercheurs les outils nécessaires pour privilégier des stratégies algorithmiques peu coûteuses en énergie, atténuant ainsi l'empreinte environnementale croissante de l'informatique.*

MOTS-CLÉS : *Graphes temporels - IA frugale - Apprentissage automatique - Partitionnement*

KEYWORDS: *Temporal Graphs - Green AI - Machine Learning - Clustering*

1. Introduction

Les données structurées en graphes sont devenues le standard pour représenter des systèmes complexes comme les réseaux de transport, les écosystèmes de médias sociaux ou les cartes d'interactions biologiques (Wu *et al.*, 2021). Bien que l'analyse de graphes statiques ait fourni des connaissances significatives, l'intégration de la dimension temporelle s'est avérée essentielle pour capturer les évolutions de ces systèmes (Rossi *et al.*, 2020). Pour exploiter ces informations spatio-temporelles, les travaux de recherche sont de plus en plus complexes, notamment **l'analyse prédictive**, et le **partitionnement**de graphes temporels(Zhang and Lei, 2023).

Afin d'évaluer et de comparer les différentes analyses, la communauté scientifique se base principalement sur les performances prédictives comme la précision. Cette quête de précision relègue au second plan l'efficacité de calculs et le coût énergétique(Schwartz *et al.*, 2020). . Cependant, les travaux liés aux graphes temporels imposent un temps d'exécution, une contrainte matérielle et une consommation énergétique considérables, en particulier lorsqu'elles sont appliquées à des jeux de données massives ou en temps continu (Luccioni *et al.*, 2023). Parallèlement, l'amélioration de l'une de ces trois dimensions, *performance prédictive*, *efficacité de calcul* ou *coût énergétique*, dégrade souvent les autres. Une plus grande précision nécessite des modèles plus complexes et donc une augmentation du temps d'exécution et de la consommation d'énergie. Inversement, des configurations économes en énergie peuvent limiter l'expressivité du modèle. Bien que des benchmarks récents aient standardisé l'évaluation de l'exactitude dans l'apprentissage sur graphes temporels (Huang *et al.*, 2023), ils manquent manifestement de métriques rigoureuses concernant l'efficacité de calcul et le coût énergétique.

De par la nature complexe des graphes temporels, nous soutenons que les performances des algorithmes d'analyses de ces graphes ne peuvent être évaluées de manière pertinente sans considérer conjointement la *performance*, l'*efficacité* et le *coût énergétique*. Contrairement aux protocoles d'évaluation existants qui se concentrent uniquement sur les métriques (F1 score, MAE, MSE...), nous proposons un cadre d'évaluation qui corrèle ces mesures standards avec des profils fins de consommation énergétique et de temps d'exécution. En analysant systématiquement la manière dont les configurations matérielles affectent le temps d'exécution, le coût énergétique et la précision prédictive, ces travaux proposent une mesure d'évaluation reposant sur un trade-off entre efficacité, performance et coût énergétique, applicable à tout algorithme évaluable en termes de performance.Ici, nous les appliquons à l'analyse prédictive et au partitionnement

2. Etat de l'Art

Cette section dresse un panorama des tâches de prédiction et de partitionnement. Ces travaux sont essentiellement basés sur la performance, mais n'apportent que très peu de réponses aux défis d'une IA éco-responsable (Green AI).

2.1. *Analyse prédictive sur graphes temporels et benchmarks*

Le domaine de l'apprentissage sur graphes temporels a évolué, passant de simples instantanés statiques à une modélisation sophistiquée en temps continu. Pour accompagner cette évolution, des bibliothèques telles que **PyTorch Geometric Temporal** (Rozemberczki *et al.*, 2021) ont standardisé l'implémentation des réseaux de neurones sur graphes dynamiques (TGNN), facilitant le déploiement d'architectures qui exploitent l'évolution et les mécanismes d'attention. En complément de ces frameworks d'apprentissage profond, des bibliothèques standards telles que Scikit Learn[1] demeurent la référence pour l'implémentation d'algorithmes de partitionnement couramment utilisés, tandis que des approches spécialisées pour le partitionnement spectral temporel ont été proposées par (Klus and Djurdjevac Conrad, 2023; Klus and Trower, 2024).

À mesure que la complexité des modèles dans ces domaines augmente, une évaluation standardisée est devenue cruciale. L'introduction récente de benchmarks pour graphes temporels (Huang *et al.*, 2023; Gastinger *et al.*, 2024) a marqué une étape importante, éloignant le domaine des jeux de données à petite échelle. Bien que ces benchmarks représentent une avancée majeure vers la comparabilité, ils demeurent fondamentalement centrés sur la précision. L'efficacité et le coût énergétique ne sont pas considérés comme des dimensions d'évaluation de premier plan.

2.2. *Compromis entre performances et efficacité dans les systèmes d'apprentissage automatique*

La tension entre les performances prédictives et l'efficacité de calcul a été largement documentée en apprentissage automatique. Canziani et al. (Canziani *et al.*, 2017) ont démontré que des gains marginaux d'exactitude dans les réseaux de neurones profonds entraînent souvent des augmentations disproportionnées du coût de calcul. Garcia-Martin et al. (García-Martín *et al.*, 2019) ont par ailleurs montré que les métriques d'efficacité, telles que l'énergie par inférence, sont essentielles pour comprendre le comportement des systèmes en conditions réelles.

Pour l'apprentissage automatique sur graphes, les premiers travaux ont commencé à caractériser l'impact calculatoire des GNN (réseaux de neurones sur graphes) en fonction du matériel, ici des GPU. GNNMark (Baruah *et al.*, 2021) a analysé l'utilisation des GPU et les goulots d'étranglement des performances dans les GNN statiques, révélant des inefficacités significatives liées à des schémas d'accès mémoire irréguliers. Cependant, ces études ne s'étendent pas aux graphes temporels, où la topologie dynamique et les mises à jour d'état dépendantes du temps introduisent une complexité supplémentaire.

1. https://scikit-learn.org/stable/

2.3. *Eco-responsabilité et performance*

Le mouvement de l'"IA verte" (*Green AI*) appelle explicitement à ce que l'efficacité énergétique et l'impact environnemental soient traités comme des critères d'évaluation de premier plan plutôt que comme des considérations secondaires (Schwartz *et al.*, 2020). Quelques travaux ont proposé de mesurer la consommation d'énergie et les émissions de carbone conjointement à l'exactitude pour les charges de travail d'apprentissage profond (Lacoste *et al.*, 2019; Luccioni *et al.*, 2023). Néanmoins, des études récentes indiquent que la précision de la mesure de la consommation énergétique peut varier de manière significative selon les charges de travail (workload)(Fischer and et al., 2025).

De plus, des travaux récents sur les systèmes d'apprentissage automatique ont souligné que les performances prédictives, l'efficacité de calcul et la consommation énergétique sont intrinsèquement couplées et ne peuvent être optimisées de manière indépendante. Des initiatives d'évaluation à grande échelle telles que MLPerf Power (Tschand *et al.*, 2025) démontrent que la réduction du temps d'exécution par l'augmentation de la capacité matérielle augmente souvent l'énergie totale consommée. Des études empiriques montrent en outre que des améliorations marginales de l'exactitude entraînent fréquemment des augmentations disproportionnées de la consommation d'énergie. Par exemple, les analyses de l'apprentissage ensembliste et de l'optimisation des hyperparamètres rapportent que les gains en termes de performances prédictives sont inversement proportionnels à l'augmentation du coût énergétique(Omar *et al.*, 2024; Ariyanti *et al.*, 2025) ; cela souligne le rôle critiquedes choix des configurations matérielles dans la définition de ces compromis.

En dépit de ces avancées, l'analyse des compromis reste largement absente de l'apprentissage sur graphes temporels. Les évaluations existantes continuent de prioriser la précision, tandis que les mesures énergétiques sont souvent rapportées sous forme de statistiques auxiliaires plutôt que comme faisant partie intégrante d'un cadre de décision structuré équilibrant *coût énergétique*, *efficacité* et *performances*.

3. Evaluation éco-responsable : un compromis entre efficacité, performances et coût énergétique

Dans cette section, nous proposons un cadre pour identifier les configurations optimales permettant d'exécuter efficacement un algorithme basé sur les graphes temporels, en équilibrant les objectifs d'efficacité, de réduction du coût énergétique et de performances.

Soit un ensemble d'algorithmes d'analyse des graphes temporels $A = \{a_1, a_2, \ldots, a_n\}$ et un jeu de données DS, nous cherchons à caractériser le profil d'exécution de A sur DS dans un environnement offrant de multiples options de configurations matérielles $C = \{c_1, c_2, \ldots, c_m\}$. Nous définissons le *score de coût global* (m_{global}) comme une somme pondérée des coûts normalisés pour $a_i \in A$, $c_j \in C$ et DS :

$$m_{\text{global}} = \alpha \times E_{\text{norm}} + \beta \times S_{\text{norm}} + \gamma \times P_{\text{norm}} \quad [1]$$

où :

– α, β, γ sont des coefficients de pondération définis par l'utilisateur tels que $\alpha + \beta + \gamma = 1$. Ces poids reflètent la priorisation des contraintes.

– E_{norm} est le temps d'exécution normalisé, **mesurant l'efficacité**.

– S_{norm} est la consommation d'énergie normalisée, **mesurant le coût énergétique**.

– P_{norm} est la *pénalité de performances* (ou erreur) normalisée, **mesurant la performance**.

Nous définissons E_{norm} et S_{norm} de la manière suivante :

$$E_{\text{norm}} = \frac{E - E_{\text{min}}}{E_{\text{max}} - E_{\text{min}}} \qquad \text{and} \qquad S_{\text{norm}} = \frac{S - S_{\text{min}}}{S_{\text{max}} - S_{\text{min}}}$$

où :

– E est le temps d'exécution de l'algorithme $a_i \in A$, avec une configuration $c_j \in C$.

– E_{min} et E_{max} sont les temps d'exécution minimal et maximal observés pour $a_i \in A, \forall c \in C$.

– S est la consommation d'énergie de l'algorithme $a_i \in A$, avec une configuration $c_j \in C$.

– S_{min} et S_{max} sont les valeurs de consommation d'énergie minimale et maximale observées pour $a_i \in A, \forall c \in C$.

Étant donné que les métriques prédictives varient selon la tâche (par ex., une valeur plus élevée est préférable pour la précision/score F1, une valeur plus faible est préférable pour la RMSE/MAE), P_{norm} est adapté pour garantir un objectif de minimisation cohérent :

– **Pour les métriques d'erreur (par ex., RMSE, MAE) :** puisque l'erreur se minimise quand sa valeur se rapproche de 0, nous normalisons :

$$P_{\text{norm}} = \frac{P - P_{\text{min}}}{P_{\text{max}} - P_{\text{min}}}$$

– **Pour les métriques de précision (par ex., score F1, AUC) :** puisque la précision se maximise quand sa valeur s'éloigne de 0, nous inversons le score normalisé pour représenter la « distance par rapport aux performances optimales » :

$$P_{\text{norm}} = 1 - \left(\frac{P - P_{\text{min}}}{P_{\text{max}} - P_{\text{min}}} \right)$$

Où :

– P sont les performances de $a_i \in A$ pour DS.

– P_{min} et P_{max} sont les valeurs de performances la plus basse et la plus haute observées, $\forall a \in A$ pour DS.

4. Configuration expérimentale

Pour caractériser empiriquement les compromis entre les performances prédictives, l'efficacité de calcul et le coût énergétique, nous avons conçu une configuration expérimentale couvrant deux familles d'algorithmes d'analyse de graphes : les TGNN et les algorithmes de partitionnement de graphes temporels. Pour des raisons de lisibilité, nous rapportons des résultats représentatifs dans cette section. Le code utilisé pour les expérimentations est disponible sur Github[2].

Jeux de données : Nous avons sélectionné 4 jeux de données spatio-temporels réels variant en taille de graphe (V) et en durée temporelle (T). Cela nous permet d'évaluer si les compromis observés sont spécifiques aux jeux de données ou structurellement induits par des facteurs algorithmiques et matériels. Leurs propriétés sont résumées dans le Tableau 1.

Nom	Description	V	T
Wikipedia Math	Nombre de vues quotidiennes d'articles Wikipédia liés aux mathématiques.	731	1068
Chickenpox Hungary	Nombre hebdomadaire de cas de varicelle recensés dans les comtés hongrois.	20	522
Windmill	Production horaire d'énergie générée par des éoliennes.	319	17472
METR-LA	Relevés de la vitesse moyenne du trafic capturés par des capteurs routiers à Los Angeles.	208	34272

TABLEAU 1. *Propriétés des jeux de données spatio-temporels utilisés dans cet article. **V** désigne le nombre de sommets et **T** désigne le nombre d'instantanés temporels (snapshots).*

Suivi énergétique : Afin de mener ces expériences, nous avons choisi **PyJoules**[3] pour suivre la consommation énergétique du CPU et du GPU des différents algorithmes. Cette bibliothèque Python permet de mesurer la consommation énergétique pour une partie spécifique d'un programme Python en utilisant pour un CPU Intel l'outil RAPL (*Running Average Power Limit*) et la bibliothèque de gestion NVIDIA (*NVIDIA Management Library*) disponible sur les GPU NVIDIA dotés de l'*architecture Volta*.

Configuration matérielle : Toutes les expériences ont été menées sur une plateforme matérielle unique pour éliminer la variabilité inter-plateformes concernant le

2. https://github.com/PPCavallera/TemporalGraphAnalysis-Consumption
3. https://pyjoules.readthedocs.io

temps d'exécution et la consommation d'énergie. Plus précisément, nous avons exécuté ces expériences à l'aide d'un ordinateur Dell Precision 5490 équipé d'un processeur Intel Core Ultra 7 165H (22 threads), d'un GPU NVIDIA RTX 2000 Ada Generation Laptop avec 8 Go de VRAM, et de 64 Go de RAM.

5. Évaluation expérimentale de l'apprentissage sur graphes temporels

Dans cette section, nous présentons les protocoles d'expérimentation et d'évaluation du compromis multi-critères pour l'apprentissage dans les graphes temporels. Dans un premier temps, nous présentons les modèles, les métriques utilisées pour analyser les performances et le protocole suivi pour mener ces expériences. Dans un second temps, nous présentons les résultats expérimentaux et analysons les compromis entre performances prédictives, efficacité et coût énergétique.

5.1. *Protocoles d'expérimentation et d'évaluation pour l'apprentissage sur graphes temporels*

Modèles sélectionnés : Nous avons sélectionné différents types d'algorithmes classés parmi les TGNN. Ces algorithmes ont pour but de prédire les valeurs des attributs des sommets dans un graphe temporel. Nous avons utilisé l'implémentation de ces modèles issue de *Pytorch Geometric Temporal* (Rozemberczki *et al.*, 2021).

– **DCRNN (Diffusion Convolutional Recurrent Neural Network)** (Li *et al.*, 2018) : Une architecture de prédiction lourde qui intègre des marches aléatoires bidirectionnelles avec un encodeur-décodeur GRU. Elle sert de référence pour les opérations coûteuses en mémoire.

– **GConvGRU (Graph Convolutional GRU)** (Seo *et al.*, 2018) : Une base de référence récurrente simplifiée qui approxime la convolution de graphe à l'aide de polynômes de Tchebychev, offrant une alternative légère au DCRNN.

– **DyGrAE (Dynamic Graph Autoencoder)** (Taheri *et al.*, 2019; Taheri and Berger-Wolf, 2019) : Une architecture non supervisée composée d'un encodeur et d'un décodeur basés sur les graphes, se concentrant sur la reconstruction d'instantanés de graphes dynamiques.

Métriques d'évaluation : La prédiction est évaluée par des métriques de régression standard. Ces métriques présentent des sensibilités différentes à l'ampleur des erreurs, offrant ainsi une évaluation complète des performances du modèle. Nous avons utilisé **l'erreur quadratique moyenne (MSE)**, **la racine de l'erreur quadratique moyenne (RMSE)** et **l'erreur absolue moyenne (MAE)**.

5.1.1. *Protocole expérimental*

Afin de quantifier rigoureusement les compromis entre performances prédictives, coût énergétique et efficacité, nous avons conçu le protocole suivant.

– **Adaptation des paramètres de l'horloge du GPU :** Pour évaluer la manière dont les architectures d'apprentissage profond réagissent aux contraintes matérielles, nous avons mis en œuvre une stratégie de plafonnement de la fréquence sur le GPU à l'aide de l'interface `nvidia-smi`. Nous avons testé 10 configurations allant de la fréquence durable minimale ($f_{min} = 495MHz$) à la fréquence d'horloge maximale ($f_{max} = 3105MHz$). Cette stratégie de mise à l'échelle de la fréquence permet d'explorer les compromis non linéaires entre le temps de résolution et la consommation d'énergie, qui ne sont pas observables avec l'ordonnancement par défaut du GPU.

– **Référentiel d'apprentissage commun :** Chaque modèle TGNN a été entraîné sur les jeux de données sélectionnés avec un budget strict de 50 époques. Ce nombre fixe élimine la variance liée à l'arrêt précoce (*early stopping*), concentrant ainsi l'évaluation sur le coût de calcul d'un processus d'entraînement.

– **Profilage des performances :** Nous avons surveillé le temps d'exécution total (TET) et le coût énergétique (EC) en temps réel. L'énergie a été mesurée en joules, en agrégeant la consommation électrique **du GPU et du CPU**. Nous considérons que les performances restent constantes en fonction de la configuration matérielle.

5.2. *Résultats expérimentaux de l'apprentissage sur graphes temporels*

Cette section présente et analyse les résultats expérimentaux obtenus lors de l'évaluation des modèles TGNN. Pour des raisons de lisibilité, seuls les résultats sur le jeu de données **Wikipedia Math** sont présentés. L'intégralité des résultats est disponible sur Github2, et présentent des comportements similaires sur les autres jeux de données. Nous avons mené ces expériences avec 10 configurations de GPU, afin d'observer (i) le comportement de chaque modèle en termes de performances prédictives, d'efficacité et de coût énergétique, et (ii) les compromis entre ces trois critères.

5.2.1. *Résultats expérimentaux concernant l'efficacité, le coût énergétique et les performances*

Nous rapportons le temps d'exécution et la consommation d'énergie sous l'effet de l'ajustement de la cadence d'horloge du GPU (cf. Figure 1), ainsi que les performances prédictives correspondantes pour chaque modèle (cf. Tableau 2).

	MSE	**MAE**	**RMSE**
DCRNN	**0.000809**	0.01320	**0.02845**
GConvGRU	0.001014	0.01424	0.03184
DyGrAE	0.000931	**0.01313**	0.03051

TABLEAU 2. *MSE, MAE et RMSE pour la prédiction sur Wikipedia Math en utilisant DCRNN, GConvGRU et DyGrAE. Nous avons mis en évidence les meilleures performances.*

Le choix du modèle comme levier majeur d'efficacité : L'analyse de la Figure 1 révèle que DyGrAE s'impose comme le modèle le plus efficient. Indépendamment de

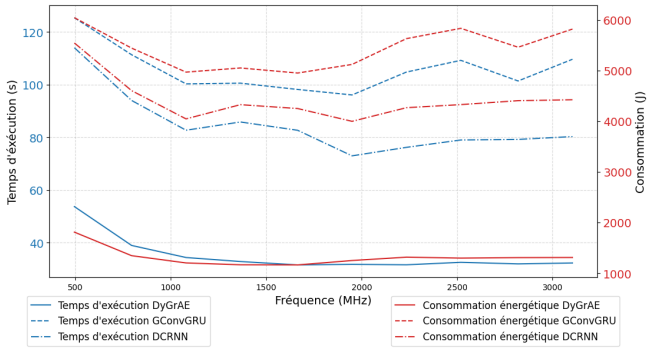


FIGURE 1. *Évolution de la consommation énergétique et du temps d'exécution des TGNN sur le jeu de données Wikipedia Math en fonction de la fréquence de GPU disponible.*

la fréquence choisie, il surpasse ses concurrents en rapidité et en sobriété énergétique. À titre d'exemple, au seuil de 1665MHz, DyGrAE complète son entrainement en seulement 32s pour une consommation de 1163J. En comparaison, le DCRNN (83s / 4250J) et GConvGRU (98s / 4950J) s'avèrent 2,5 à 3 fois plus lents et jusqu'à 4 fois plus énergivores. De plus, comme présenté dans le Tableau 2, il présente une précision compétitive, se révélant même supérieur aux autres modèles dans le cas de la MAE (0,01313).

Non-linéarité entre fréquence et performance énergétique : Dans la Figure 1, nous remarquons comme attendu une corrélation forte entre temps de calcul et consommation énergétique. Nous pouvons cependant noter que cette relation se déteriore pour des fréquences plus hautes ($> 2000MHz$). Ces résultats démontrent aussi que la performance optimale ne coïncide pas avec la fréquence maximale du GPU. On observe pour chaque architecture une zone de compromis "énergie-temps" située à des fréquences intermédiaires. DCRNN atteint son point d'inflexion optimal autour de 1950MHz (73s, 3997J). Par rapport à une fréquence basse de 495,MHz, ce réglage permet de réduire le temps d'exécution de 36,0% et la consommation de 27,9%. Paradoxalement, pousser la fréquence au maximum (3105,MHz) dégrade le bilan : la consommation remonte à 4423J (+10,7%) et le temps d'exécution à 80s (+9,6%). DyGrAE et GConvGRU suivent des schémas similaires, avec une efficacité énergétique maximale identifiée à 1665MHz (respectivement 1163J et 4950J). Ces résultats soulignent qu'au-delà d'un certain seuil (environ 2000MHz ici), l'augmentation de la cadence matérielle engendre, logiquement, une augmentation de la consommation énergétique, mais n'aura qu'un impact mineur sur le temps d'exécution, voire dégradera le temps d'entrainement.

			DyGrAE		GConvGRU		DCRNN	
α	β	γ	**Compromis**	**Fréq GPU**	**Compromis**	**Fréq GPU**	**Compromis**	**Fréq GPU**
1	**0**	**0**	**0**	**1665**	0.688	1950	0.441	1950
0	**1**	**0**	**0**	**1665**	0.776	1665	0.581	1950
0	**0**	**1**	0.607	_	1	_	**0**	_
0	**0.5**	**0.5**	0.303	1665	0.888	1665	**0.29**	**1950**
0.5	**0**	**0.5**	0.303	1665	0.844	1950	**0.221**	**1950**
0.5	**0.5**	**0**	**0**	**1665**	0.743	1665	0.511	1950
0.33	**0.33**	**0.33**	**0.2**	**1665**	0.82	1665	0.337	1950
0.2	**0.2**	**0.6**	0.364	1665	0.897	1665	**0.204**	**1950**

TABLEAU 3. *Évolution du compromis en fonction des valeurs de α, β, et γ pour **DCRNN, DyGrAE, GConvGRU**. Ce tableau présente pour chaque modèle la meilleure valeur de compromis et la fréquence de GPU associée (**Fréq GPU**) en MHz. La valeur de fréquence pour un compromis axé uniquement sur les performances est non pertinente car nous la considérons comme constante indépendamment des configurations de GPU.*

5.2.2. *Compromis multi-critères appliqué à l'apprentissage sur graphes temporels*

Nous examinons les différents modèles TGNN et analysons le compromis entre efficacité, coût énergétique et performances en fonction de la fréquence de GPU disponible. Nous présentons dans le Tableau 3 la meilleure valeur de compromis et la meilleure configuration matérielle pour différentes combinaisons de α, β, et γ.

Dans le Tableau 3, nous observons une distinction nette dans la spécialisation des modèles selon les objectifs visés. DyGrAE s'impose systématiquement comme la solution la plus robuste pour les compromis privilégiant le coût énergétique, l'efficacité, ou une combinaison équilibrée des deux ($\alpha, \beta \geq 0.5$). Cette prédominance confirme les observations de la Figure 1, soulignant sa capacité à maintenir une faible empreinte énergétique sans sacrifier la qualité du traitement. À l'inverse, DCRNN démontre une meilleure aptitude lorsque la priorité est accordée à la dimension des performances pures ($\gamma = 1$). Sur le plan matériel, les configurations de GPU optimales ne se situent jamais aux extrêmes de la plage de fréquences, mais gravitent entre 1665 MHz et 1950 MHz. Cette plage, représentant environ 60 % de la fréquence maximale, elle offre une puissance de calcul suffisante pour éviter l'allongement excessif du temps d'exécution, préjudiciable au coût énergétique, tout en évitant la consommation élevées nécessaire aux hautes fréquences. Ainsi, l'ajustement dynamique de la cadence d'horloge apparaît comme un levier d'optimisation nécessaire pour adapter l'infrastructure à la réduction de la consomamtion énergétique des algorithmes prédictifs.

6. Évaluation expérimentale du partitionnement de graphes temporels

Dans cette section, nous présentons les protocoles d'expérimentation et d'évaluation du compromis multi-critères pour les algorithmes de partitionnement de graphes temporels. Dans un premier temps, nous présentons donc les modèles, les métriques utilisées pour analyser les performances et le protocole suivi pour mener ces expériences.

Dans un second temps, nous présentons les résultats expérimentaux et analysons les compromis entre qualité du partitionnement, efficacité et coût énergétique.

6.1. *Protocoles d'expérimentation et d'évaluation pour le partitionnement de graphes temporels*

Les algorithmes de partitionnement offrent une perspective complémentaire aux TGNN, car ils reposent sur des schémas de calcul de natures différentes.

Modèles sélectionnés : Nous avons sélectionné un ensemble d'algorithmes de partitionnement (algorithmes courants et algorithmes de partitionnement spécifiques aux graphes temporels afin d'évaluer différents types d'approches).

– **K-Means** : Une méthode de partitionnement basée sur la répartition qui minimise la variance au sein de k clusters. Elle sert de référence efficace sur le plan des calculs. L'implémentation de l'algorithme est tirée de Scikit Learn[1].

– **Partitionnement agglomératif (*Agglomerative Clustering*)** : Une approche hiérarchique qui fusionne des paires de clusters en fonction d'une métrique distance donnée. L'implémentation de l'algorithme est tirée de Scikit Learn[1].

– **Partitionnement spectral (*Spectral Clustering*)** : Une technique qui utilise les valeurs propres de la matrice d'adjacence pour effectuer une réduction de dimensionnalité avant le partitionnement. L'implémentation de l'algorithme est tirée l'article (Klus and Djurdjevac Conrad, 2023).

Qualité du partitionnement : Nous utilisons des métriques non-supervisées. Ces métriques visent à évaluer la qualité du partitionnement selon différentes approches.

– **Score de Davies-Bouldin (DB) (Davies and Bouldin, 1979) :** Mesure le ratio de similarité moyen de chaque cluster avec son cluster le plus similaire. Des scores plus faibles indiquent une meilleure séparation.

– **Score de Calinski-Harabasz (CH) (Calinski and Harabasz, 1974) :** Ratio entre la somme de la dispersion inter-clusters et la dispersion intra-cluster. Des scores plus élevés indiquent des clusters denses et bien séparés.

– **Entropie moyenne des attributs (MFE) (Dash *et al.*, 1997) :** Évalue le caractère aléatoire des distributions de attributs au sein des clusters. Une entropie plus faible implique des clusters plus homogènes.

Protocole expérimental : Afin de quantifier rigoureusement les compromis entre performances de partitionnement et impact environnemental, nous avons conçu le protocole suivant, adapté à la nature calculatoire de chaque classe d'algorithmes.

– **Analyse du passage à l'échelle matérielle (CPU) :** Comme ces algorithmes d'apprentissage automatique classiques reposent principalement sur des calculs CPU, nous avons évalué leur passage à l'échelle en faisant varier le nombre de cœurs physiques disponibles. Nous avons échelonné l'allocation de 1 cœur à 22 cœurs. Cela nous permet d'identifier le point de saturation où l'ajout de cœurs supplémentaires

offre des rendements décroissants en termes de temps d'exécution par rapport au coût énergétique (Loi d'Amdahl) (Amdahl, 1967).

– **Budget d'entraînement standardisé :** Chaque modèle d'algorithme a été entraîné sur les jeux de données sélectionnés 100 fois. Comme les algorithmes de partitionnement sont rapides, exécuter le modèle plusieurs fois évite la variance provenant de potentielles latences de l'ordinateur, et par conséquent, des différences importantes de temps d'exécution.

– **Profilage des performances :** De manière similaire au protocole des TGNN, nous avons surveillé le temps d'exécution total (TET) et la coût énergétique (EC). Cependant, pour ce protocole, la mesure de l'énergie a été strictement limitée aux domaines d'alimentation du CPU, puisqu'aucune ressource GPU n'a été utilisée. Nous calculons le temps d'exécution et la consommation d'énergie comme la moyenne des 100 itérations. Nous considérons que les performances de l'algorithme restent constantes quelle que soit la configuration matérielle.

6.2. *Évaluation expérimentale du partitionnement de graphes temporels*

Cette section présente et analyse les résultats expérimentaux obtenus lors de l'évaluation des algorithmes de partitionnement. Pour des raisons de lisibilité, seuls les résultats sur le jeu de données **Wikipedia Math** sont présentés. L'intégralité des résultats est disponibles sur Github2, et présentent des comportements similaires sur les autres jeux de données. Nous avons mené ces expériences avec 22 configurations de CPU, afin d'observer (i) premièrement le comportement de chaque modèle en termes de qualité de partitionnement, d'efficacité et decoût énergétique, (ii) deuxièmement les compromis entre ces trois critères.

6.2.1. *Résultats expérimentaux concernant l'efficacité, le coût énergétique et les performances*

	DB	**CH**	**MFE**
KMeans	4.01	64.24	**4.63**
Partitionnement agglomératif	**1.39**	**185.49**	4.76
Partitionnement spectral	10.26	14.85	4.89

TABLEAU 4. *Score de Davies-Bouldin (DB), Score de Calinski-Harabasz (CH) et Entropie moyenne des attributs (MFE) pour le partitionnement de Wikipedia Math en utilisant KMeans, le partitionnement agglomératif et le partitionnement spectral.*

Qualité du partitionnement. Dans le Tableau 4, nous observons que le partitionnement agglomératif surpasse nettement les autres méthodes avec un DB score de 1,39 et un indice CH de 185,49. KMeans se distingue toutefois par une meilleure homogénéité des caractéristiques (MFE de 4,63). Les performances en retrait du partitionnement spectral (DB de 10,26) s'expliquent par le fait que l'évaluation repose sur les attributs des sommets, alors que cet algorithme privilégie la topologie structurelle du graphe.

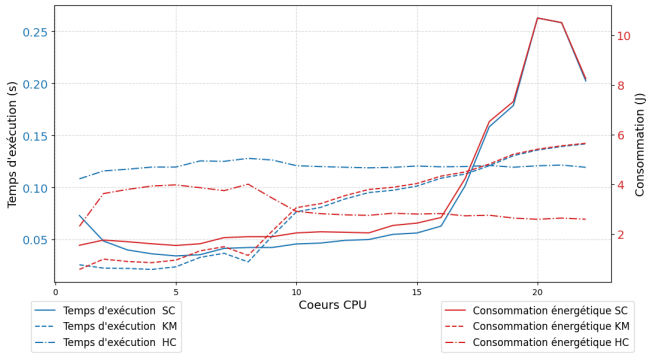


FIGURE 2. *Évolution de la consommation d'énergie et du temps d'exécution pour les algorithmes de partitionnement sur le jeu de données Wikipedia Math en fonction du nombre de cœurs CPU.*

Efficacité et coût énergétique en fonction de la puissance disponible du CPU. La Figure 2 confirme que la consommation énergétique du CPU est étroitement corrélée au temps d'exécution. L'ajout de cœurs CPU s'avère contre-productif pour KMeans et le partitionnement spectral : leurs temps d'exécution respectifs bondissent de 0,02s et 0,07s à 1 cœur vers des pics de 0,14s et 0,2s à 20 cœurs, entraînant une hausse de l'énergie consommée de près de 500% pour le spectral (passant de 1,8J à 10,5J). À l'inverse, le partitionnement agglomératif affiche une plus grande sobriété, sa consommation restant stable autour de 3J après le seuil de 10 cœurs, soulignant une saturation précoce de ses besoins en ressources.

6.2.2. *Compromis multi-critères appliqué aux algorithmes de partitionnement*

			Partitionnement spectral		**KMeans**		**Partitionnement hiérarchique**	
α	β	γ	**Compromis**	**Cœurs CPU**	**Compromis**	**Cœurs CPU**	**Compromis**	**Cœurs CPU**
1	**0**	**0**	0.054	5	**0**	**4**	0.63	1
0	**1**	**0**	0.095	5	**0**	**1**	0.172	1
0	**0**	**1**	1	_	0.268	_	**0**	_
0	**0.5**	**0.5**	0.548	5	0.134	1	**0.086**	**1**
0.5	**0**	**0.5**	0.527	5	**0.134**	**4**	0.18	1
0.5	**0.5**	**0**	0.075	5	**0.009**	**1**	0.266	1
0.33	**0.33**	**0.33**	0.379	5	**0.095**	**1**	0.176	1
0.6	**0.2**	**0.2**	0.63	5	0.165	1	**0.106**	**1**

TABLEAU 5. *Évolution du compromis en fonction des valeurs de α, β, et γ pour le **Partitionnement spectral, Partitionnement hiérarchique, KMeans**. Ce tableau présente pour chaque algorithme la meilleure valeur de compromis et le nombre de cœurs associé.*

L'analyse des résultats présentés dans le Tableau 5 montre que KMeans se révèle particulièrement adapté aux compromis privilégiant l'efficacité et le coût énergétique,

présentant le score de 0 dans les configurations ne comprenant que α, β ou les deux. À l'inverse, le partitionnement hiérarchique domine lorsque l'accent est mis sur la performance, surpassant le partitionnement spectral et KMeans. Le partitionnement spectral semble globalement moins compétitif, ses scores restant élevés (ex : 0,379 pour un compromis neutre) face aux résultats de KMeans (0,095). Sur le plan matériel, le nombre optimal de cœurs CPU demeure restreint, se stabilisant entre 1 et 5 cœurs. Cette configuration ne mobilise que 5% à 25% des ressources disponibles, suggérant qu'une parallélisation massive est contre-productive. On note d'ailleurs que KMeans et le partitionnement hiérarchique atteignent leurs meilleurs compromis avec seulement 1 cœur dans la majorité des scénarios multi-critères, confirmant qu'une sobriété matérielle tend à favoriser l'équilibre global.

7. Conclusion et Futurs Travaux

Cette étude établit un changement de paradigme nécessaire dans l'analyse de graphes temporels, dépassant les évaluations centrées sur la précision pour adopter une vision qui intègre le coût énergétique et l'efficacité. Notre cadre multi-critères révèle que le modèle optimal est fortement dépendant du contexte. En quantifiant ces trois dimensions, nous donnons aux chercheurs les moyens de prendre des décisions éclairées et respectueuses de l'environnement. Sur les plus de 13.000 tests effectués, les résultats obtenus soulignent que l'augmentation des ressources matérielles, qu'il s'agisse de la fréquence GPU ou du nombre de cœurs CPU, n'est pas systématiquement corrélée à un gain d'efficacité. Au contraire, une gestion sobre des ressources permet souvent d'atteindre un point d'équilibre optimal entre coût énergétique, temps de calcul et qualité prédictive. Les travaux futurs étendront nos travaux à des benchmarks à plus grande échelle tels que TGB (Gastinger *et al.*, 2024), à différentes architectures matérielles. À terme, nous visons à exploiter ces profils d'efficacité pour développer un cadre de méta-apprentissage capable de recommander automatiquement l'architecture la plus durable, adaptée aux besoins de l'utilisateur, ancrant ainsi l'analyse de graphes dans une démarche pérenne écologiquement.

Bibliographie

Amdahl G. M., « Validity of the single processor approach to achieving large scale computing capabilities », *Proceedings of the April 18-20, 1967, spring joint computer conference on - AFIPS '67 (Spring)*, ACM Press, Atlantic City, New Jersey, p. 483, 1967.

Ariyanti S., Suryanegara M., Arifin A. S., Nurwidya A. I., Hayati N., « Trade-Off Between Energy Consumption and Three Configuration Parameters in Artificial Intelligence (AI) Training : Lessons for Environmental Policy », *Sustainability*, June, 2025. Company : Multidisciplinary Digital Publishing Institute Distributor : Multidisciplinary Digital Publishing Institute Institution : Multidisciplinary Digital Publishing Institute Label : Multidisciplinary Digital Publishing Institute Publisher : publisher.

Baruah T., Shivdikar K., Dong S., Sun Y., Mojumder S. A., Jung K., Abellan J. L., Ukidave Y., Joshi A., Kim J., Kaeli D., « GNNMark : A Benchmark Suite to Characterize Graph

Neural Network Training on GPUs », *2021 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, IEEE, Stony Brook, NY, USA, p. 13-23, March, 2021.

Calinski T., Harabasz J., « A dendrite method for cluster analysis », *Communications in Statistics - Theory and Methods*, vol. 3, nº 1, p. 1-27, 1974.

Canziani A., Paszke A., Culurciello E., « An Analysis of Deep Neural Network Models for Practical Applications », April, 2017. arXiv :1605.07678 [cs].

Dash M., Liu H., Yao J., « Dimensionality reduction of unsupervised data », *Proceedings Ninth IEEE International Conference on Tools with Artificial Intelligence*, IEEE Comput. Soc, Newport Beach, CA, USA, p. 532-539, 1997.

Davies D. L., Bouldin D. W., « A Cluster Separation Measure », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, nº 2, p. 224-227, April, 1979.

Fischer R., et al., « Ground-Truthing AI Energy Consumption : Validating CodeCarbon Against External Measurements », *arXiv preprint*, 2025.

García-Martín E., Rodrigues C. F., Riley G., Grahn H., « Estimation of energy consumption in machine learning », *Journal of Parallel and Distributed Computing*, vol. 134, p. 75-88, December, 2019.

Gastinger J., Huang S., Galkin M., Loghmani E., Parviz A., Poursafaei F., Danovitch J., Rossi E., Koutis I., Stuckenschmidt H., Rabbany R., Rabusseau G., « TGB 2.0 : A Benchmark for Learning on Temporal Knowledge Graphs and Heterogeneous Graphs », 2024.

Huang S., Poursafaei F., Danovitch J., Fey M., Hu W., Rossi E., Leskovec J., Bronstein M., Rabusseau G., Rabbany R., « Temporal Graph Benchmark for Machine Learning on Temporal Graphs », 2023.

Klus S., Djurdjevac Conrad N., « Koopman-Based Spectral Clustering of Directed and Time-Evolving Graphs », *Journal of Nonlinear Science*, vol. 33, nº 1, p. 8, February, 2023.

Klus S., Trower M., « Transfer operators on graphs : spectral clustering and beyond », , vol. 5, nº 1, p. 015014, 2024.

Lacoste A., Luccioni A., Schmidt V., Dandres T., « Quantifying the Carbon Emissions of Machine Learning », *arXiv preprint*, 2019.

Li Y., Yu R., Shahabi C., Liu Y., « Diffusion Convolutional Recurrent Neural Network : Data-Driven Traffic Forecasting », 2018.

Luccioni A. S., Luccioni S., Co H., Viguier S., Ai G., Ligozat A.-L., Ligozat A.-L., « Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model », 2023.

Omar R., Bogner J., Muccini H., Lago P., Martínez-Fernández S., Franch X., « The More the Merrier ? Navigating Accuracy vs. Energy Efficiency Design Trade-Offs in Ensemble Learning Systems », July, 2024. arXiv :2407.02914 [cs].

Rossi E., Chamberlain B., Frasca F., Eynard D., Monti F., Bronstein M., « Temporal Graph Networks for Deep Learning on Dynamic Graphs », October, 2020. arXiv :2006.10637 [cs].

Rozemberczki B., Scherer P., He Y., Panagopoulos G., Riedel A., Astefanoaei M., Kiss O., Beres F., López G., Collignon N., Sarkar R., « PyTorch Geometric Temporal : Spatiotemporal Signal Processing with Neural Machine Learning Models », June, 2021. arXiv :2104.07788 [cs].

Schwartz R., Dodge J., Smith N. A., Etzioni O., « Green AI », *Communications of the ACM*, vol. 63, nº 12, p. 54-63, November, 2020.

Seo Y., Defferrard M., Vandergheynst P., Bresson X., « Structured Sequence Modeling with Graph Convolutional Recurrent Networks », *in* L. Cheng, A. C. S. Leung, S. Ozawa (eds), *Neural Information Processing*, vol. 11301, Springer International Publishing, Cham, p. 362-373, 2018. Series Title : Lecture Notes in Computer Science.

Taheri A., Berger-Wolf T., « Predictive temporal embedding of dynamic graphs », *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ACM, Vancouver British Columbia Canada, p. 57-64, August, 2019.

Taheri A., Gimpel K., Berger-Wolf T., « Learning to Represent the Evolution of Dynamic Graphs with Recurrent Models », *Companion Proceedings of The 2019 World Wide Web Conference*, ACM, San Francisco USA, p. 301-307, May, 2019.

Tschand A., Rajan A. T. R., Idgunji S., Ghosh A., Holleman J., Kiraly C., Ambalkar P., Borkar R., Chukka R., Cockrell T., Curtis O., Fursin G., Hodak M., Kassa H., Lokhmotov A., Miskovic D., Pan Y., Manmathan M. P., Raymond L., John T. S., Suresh A., Taubitz R., Zhan S., Wasson S., Kanter D., Reddi V. J., « MLPerf Power : Benchmarking the Energy Efficiency of Machine Learning Systems from Microwatts to Megawatts for Sustainable AI », February, 2025. arXiv :2410.12032 [cs].

Wu Z., Pan S., Chen F., Long G., Zhang C., Yu P. S., « A Comprehensive Survey on Graph Neural Networks », *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, n° 1, p. 4-24, January, 2021.

Zhang C., Lei M., « A Survey on Spatio-Temporal Graph Neural Networks for Traffic Forecasting », *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, IEEE, Shanghai, China, p. 1417-1423, December, 2023.

Ingénierie d'un SI de vidéosurveillance responsable : arbitrage énergétique entre CNN 3D personnalisés et modèles vision-langage en apprentissage fédéré

Sébastien Thuau[1,2], Siba Haidar[1], Rachid Chelouah[2]

1. esieaLab, ESIEA, 75005 Paris, France
sebastien.thuau@ext.esiea.fr ; siba.haidar@esiea.fr

2. ETIS (UMR 8051), CY Cergy Paris Université, CNRS, 95000 Cergy, France
rc@cy-tech.fr

RESUME. *Cet article est une synthèse de l'article : S. Thuau, S. Haidar, R. Chelouah, « Federated Learning for Video Violence Detection: Complementary Roles of Lightweight CNNs and Vision-Language Models for Energy-Efficient Use », 2025 IEEE 37th International Conference on Tools with Artificial Intelligence (ICTAI), Athens, Greece, 2025, pp. 739–743, DOI: 10.1109/ICTAI66417.2025.00106. Ce travail examine l'intégration de modèles d'intelligence artificielle au sein d'un système d'information de vidéosurveillance distribué en considérant des contraintes non fonctionnelles majeures : confidentialité, hétérogénéité des données (non-IID), consommation énergétique et empreinte carbone. Trois stratégies sont comparées : inférence zéro-shot de modèles vision-langage, adaptation fédérée par LoRA d'un modèle multimodal, et apprentissage fédéré personnalisé d'un CNN 3D léger. Les résultats mettent en évidence un arbitrage entre capacité multimodale et sobriété énergétique. Cette synthèse en propose une lecture orientée ingénierie des systèmes d'information, centrée sur l'arbitrage multicritère performance–énergie–confidentialité et la gouvernance des modèles.*

MOTS-CLÉS : *systèmes d'information, intelligence artificielle responsable, apprentissage fédéré, confidentialité des données, efficacité énergétique, empreinte carbone, modèles multimodaux, gouvernance des SI*

1. Introduction

Les systèmes d'information de vidéosurveillance s'inscrivent aujourd'hui dans des environnements organisationnels complexes caractérisés par la distribution des données, la sensibilité des contenus et des exigences croissantes en matière de responsabilité environnementale. Leur conception implique une prise en compte explicite de contraintes non fonctionnelles telles que la confidentialité des données, l'hétérogénéité inter-organisationnelle, la consommation énergétique et l'empreinte carbone.

Dans ce contexte, l'apprentissage fédéré constitue un mécanisme structurant pour concevoir des architectures distribuées préservant la localité des données. Toutefois, l'essor des modèles multimodaux de grande taille, notamment les modèles vision-langage, introduit une tension entre capacité de raisonnement contextuel et sobriété computationnelle.

La problématique centrale devient alors architecturale : comment intégrer différents types de modèles d'intelligence artificielle dans un système d'information distribué tout en maîtrisant simultanément performance, consommation énergétique et gouvernance des mises à jour ?

La présente synthèse propose une lecture orientée ingénierie des systèmes d'information d'une étude comparative menée dans un cadre d'apprentissage fédéré. Elle vise à mettre en évidence les implications architecturales et organisationnelles des choix de modèles, au-delà des seules métriques prédictives.

2. Architecture et stratégies d'intégration de l'IA

L'architecture étudiée repose sur un système d'information distribué où les flux vidéo restent localisés au sein des entités productrices de données. L'apprentissage fédéré constitue le mécanisme d'orchestration des modèles, garantissant l'absence de transfert des données brutes tout en introduisant des contraintes de communication et de gestion de l'hétérogénéité (non-IID).

Trois stratégies d'intégration de l'intelligence artificielle sont analysées. La première mobilise des modèles vision-langage en inférence zéro-shot, privilégiant la capacité de généralisation mais au prix d'un coût computationnel élevé. La deuxième repose sur l'adaptation fédérée par LoRA d'un modèle multimodal de grande taille, réduisant les paramètres échangés tout en conservant une expressivité élevée. La troisième s'appuie sur un CNN 3D léger entraîné en apprentissage fédéré personnalisé, articulant agrégation globale et spécialisation locale dans une logique de sobriété énergétique.

Ces stratégies traduisent des choix architecturaux distincts pour un système d'information intégrant l'IA. L'enjeu ne réside pas uniquement dans la performance des modèles, mais dans leur orchestration cohérente au sein d'une architecture capable d'arbitrer entre expressivité multimodale et frugalité computationnelle.

3. Résultats et implications pour l'ingénierie des SI

Les résultats expérimentaux mettent en évidence un arbitrage structurant entre performance prédictive, qualité de calibration et coût énergétique. Si les modèles multimodaux présentent des capacités supérieures en matière de compréhension contextuelle et de classification multiclasses, leur adaptation fédérée engendre un coût énergétique sensiblement plus élevé que celui observé pour le modèle convolutionnel léger.

Ces observations doivent cependant être interprétées avec prudence. Les expérimentations ont été conduites dans un cadre simulé et sur des corpus spécifiques ; la généralisation à des infrastructures réelles implique des considérations supplémentaires telles que la variabilité matérielle, la latence réseau et la gouvernance organisationnelle des mises à jour. Cette limite renforce l'intérêt d'une approche d'ingénierie des systèmes d'information intégrant dès la conception les dimensions opérationnelles et organisationnelles du déploiement.

Le modèle 3D CNN personnalisé obtient des performances comparables en détection binaire tout en divisant approximativement par deux la consommation énergétique observée lors de l'entraînement fédéré. Cette différence, bien que mesurée dans un cadre expérimental contrôlé, devient significative à l'échelle d'un parc de dispositifs distribués.

Ces résultats confirment la pertinence d'une évaluation multicritère intégrant explicitement des métriques environnementales dans l'ingénierie des systèmes d'information.

Une architecture hybride se dégage de cette analyse : un filtrage continu assuré par un modèle frugal local, combiné à une mobilisation ponctuelle d'un modèle multimodal pour les cas ambigus ou nécessitant une interprétation contextuelle approfondie. Une telle orchestration permet de concilier efficacité opérationnelle et maîtrise des impacts énergétiques.

Du point de vue de la gouvernance des systèmes d'information, cette approche soulève plusieurs enjeux : pilotage des modèles en production, suivi des indicateurs énergétiques, gestion des mises à jour fédérées, et traçabilité des décisions algorithmiques. L'intégration de l'IA devient ainsi un problème de gouvernance technique et organisationnelle autant qu'un problème d'optimisation algorithmique.

4. Conclusion

Cette synthèse met en évidence la nécessité d'une approche architecturale intégrant explicitement contraintes énergétiques, confidentialité et hétérogénéité organisationnelle dans les systèmes d'information distribués. L'intégration de l'IA ne relève pas d'un simple choix de modèle, mais d'un arbitrage structurant entre performance, sobriété et gouvernance. La combinaison raisonnée de modèles frugaux et multimodaux constitue une piste concrète pour concevoir des systèmes d'information intelligents et responsables.

Vers une Réponse de Sécurité Autonome Guidée par une Intention et Reposant sur une Ontologie

Zequan Huang[1,2], Jacques Robin[2], Nicolas Herbaut[1], Nourhène Ben Rabah[1], Bénédicte Le Grand[1]

1. Centre de Recherche Informatique, Université Paris 1 Panthéon-Sorbonne
31 rue Baudricourt, 75013 Paris, France
{nicolas.herbaut, nourhene.ben-rabah, benedicte.le-grand}@univ-paris1.fr

2. esieaLab, École Supérieure d'Informatique Électronique Automatique
4 All. Katherine Johnson, 94200 Ivry-sur-Seine, France
{zequan.huang, jacques.robin}@esiea.fr

Cet article est une synthèse de l'article : Zequan Huang, Jacques Robin, Nicolas Herbaut, Nourhène Ben Rabah, Bénédicte Le Grand (2025). Toward an Intent-Based and Ontology-Driven Autonomic Security Response in Security Orchestration Automation and Response. International Conference on Enterprise Design, Operations, and Computing, 266-283.

1. Introduction

Les infrastructures réseaux et systèmes modernes font face à des menaces sophistiquées, notamment les menaces persistantes avancées, obligeant les plateformes pour ***l'Orchestration, l'Automatisation et la Réponse à la Sécurité (SOAR)*** à s'adapter rapidement. Pour la défense, si la prévention et la détection sont cruciales, la réponse aux incidents (post-détection) est essentielle, car un attaquant qualifié parviendra probablement à contourner les premières lignes de défense.

Actuellement, les ***Centres Opérationnels de Sécurité (SOC)*** reposent encore largement sur des experts humains pour analyser et répondre aux incidents, ce qui entraîne une fatigue face au volume de faux positifs (fausses alertes), menant à des erreurs de jugement inévitables (Zidan et al., 2024). Bien que les plateformes SOAR offrent une certaine automatisation via des *playbooks*, ceux-ci fonctionnent comme des listes de procédures rigides et de bas niveau, sans intégrer les objectifs stratégiques de haut niveau (*i.e.,* intention) des analystes SOC (Kinyua et Awuah, 2021).

Pour combler ce fossé entre les opérations procédurales et les objectifs stratégiques, nous proposons de rapprocher deux domaines de recherche : la ***Cyber Défense guidée par une Intention (ICD)*** (Leivadeas et Falkner, 2023) et la ***Cyber***

Défense Autonome (ACD) (Vyas et al., 2023). L'objectif est de permettre aux SOAR de nouvelle génération de générer et d'appliquer autonomiquement des intentions de sécurité, plutôt que d'exécuter des playbooks statiques.

Cet article présente une définition unifiée de l'intention de sécurité, fondée sur l'ontologie D3FEND (Kaloroumakis et Smith, 2021), et propose deuxièmement une architecture de ***Cyber Défense Autonome Guidée par une Intention*** pour les SOAR de nouvelle génération. Il s'agit d'un modèle à deux niveaux intégrant la découverte et l'exécution de ces intentions dans des modèles de décision théoriques, pour une réponse automatisée et robuste.

2. Définition unifiée de l'intention de sécurité

Pour permettre une défense autonome basée sur une intention, il est impératif de formaliser ce qu'est une intention de sécurité de manière exploitable par une machine. Notre approche s'appuie sur l'ontologie de cybersécurité D3FEND pour structurer les connaissances défensives. Nous privilégions cette ontologie pour son alignement natif avec la taxonomie offensive ATT&CK (Strom et al., 2017), et pour la facilité de corrélation offerte entre les menaces observées et les réponses aux incidents possibles.

Au cœur de cette formalisation se trouve le concept d'artefact numérique. Il s'agit d'une entité numérique (*e.g.,* un fichier, un processus) sur laquelle une technique offensive (issue d'ATT&CK) agit pour compromettre le système, et sur laquelle une technique défensive (issue de D3FEND) doit agir pour neutraliser cette menace.

L'objectif d'une intention est de prescrire une action défensive sur un artefact précis pour invalider les effets de l'attaque. Nous avons étendu l'ontologie D3FEND en enrichissant la sémantique des propriétés offensives (*e.g., Altérer*) et défensives (*e.g., Isoler*). Une réponse est considérée comme valide si la catégorie de l'action défensive choisie est sémantiquement compatible avec celle de l'attaque.

3. Architecture de cyber défense autonome guidée par une intention

L'intégration des intentions de sécurité dans un système de défense autonome nécessite une architecture capable de gérer à la fois la prise de décision stratégique et l'application opérationnelle. Nous proposons une approche à deux niveaux, inspirée du cycle de vie du Réseau basé sur l'Intention (Clemm et al., 2021), distinguant un ***Agent de Découverte d'Intention (IDA)*** et un ***Agent d'Exécution d'Intention (IEA)***.

3.1. Agent de Découverte d'Intention (IDA)

La fonction de l'IDA est de générer des intentions adaptées au contexte. Face à une observation de sécurité (*i.e.,* alertes), l'IDA projette les techniques offensives détectées sur l'ontologie D3FEND étendue pour identifier les contre-mesures (sous forme d'intentions) candidates. Il résout ensuite un processus de décision markovien partiellement observable (Hammar et al., 2025) pour choisir une stratégie maximisant l'utilité à long terme (*e.g.,* ajout, modification d'une intention).

3.2. Agent d'Exécution d'Intention (IEA)

L'IEA prend en charge la traduction, l'exécution et la surveillance des intentions. Il étend les capacités des cadres existants, tels que le système proposé par (Lingga et al. 2025).

4. Conclusion

Cet article propose une avancée vers des plateformes SOAR de nouvelle génération, favorisant une supervision stratégique plutôt qu'une intervention manuelle directe. En rapprochant l'ICD et l'ACD à travers une définition d'intention de sécurité pilotée par l'ontologie D3FEND, nous avons introduit une architecture de ***Cyber Défense Autonome guidée par une Intention*** à deux niveaux (IDA et IEA). Elle permet de découpler les objectifs de sécurité de leur mise en œuvre technique. Les travaux futurs se concentreront sur l'évaluation de performance des implémentations de notre proposition face à des scénarios d'attaques dynamiques et complexes.

Remerciements :

Ce travail a été soutenu par l'Agence Nationale de la Recherche (ANR) dans le cadre du projet ANCILE (n° ANR-23-CE39-0010).

Bibliographie

Clemm et al. (2021). Intent-Based Networking - Concepts and Definitions. RFC 9315, IETF.

Hammar et al. (2025). Adaptive Security Response Strategies Through Conjectural Online Learning. IEEE Transactions on Information Forensics and Security, 20, 4055-4070.

Kaloroumakis et Smith. (2021). Toward a knowledge graph of cybersecurity countermeasures. The MITRE Corporation.

Kinyua et al. (2021). AI/ML in Security Orchestration, Automation and Response: Future Research Directions. Intelligent Automation & Soft Computing, 28(2), 527-545.

Leivadeas et al. (2023). A Survey on Intent-Based Networking. IEEE Communications Surveys & Tutorials, 25(1), 625-655.

Lingga et al. (2025). ICSC: Intent-Based Closed-Loop Security Control System for Cloud-Based Security Services. IEEE Communications Magazine, 63(4), 169-175.

Strom et al. (2017). Finding cyber threats with ATT&CK-based analytics. The MITRE Corporation, Bedford, MA, Technical Report No. MTR170202.

Vyas et al. (2023). Automated cyber defence: A review. arXiv preprint arXiv:2303.04926.

Zidan et al. (2024). Assessing the Challenges Faced by Security Operations Centres (SOC). Advances in Information and Communication, 256–271.

Repenser l'Évaluation et la Classification des Ontologies sur la Cybersécurité : Vers un Cadre Centré autour de la Crédibilité

Antoine Leblanc[1,2], Jacques Robin[2], Nourhène Ben Rabah[1], Zequan Huang[1,2], Bénédicte Le Grand[1]

1. Centre de Recherche en Informatique (CRI), Université Paris 1 Panthéon Sorbonne
12 place du Panthéon, 75231 Paris cedex 05, France
{antoine.leblanc, nourhene.ben-rabah, benedicte.le-grand, zequan.huang}@univ-paris1.fr

2. esieaLab, École Supérieure d'Informatique Électronique Automatique
4 All. Katherine Johnson, 94200 Ivry-sur-Seine, France
{antoine.leblanc, jacques.robin, zequan.huang}@esiea.fr

REFERENCE DE L'ARTICLE INTERNATIONAL. Cet article est une synthèse de l'article : Leblanc, A., Robin, J., Ben Rabah, N., Huang, Z., Le Grand, B. (2026). Rethinking Cybersecurity Ontology Classification and Evaluation: Towards a Credibility-Centered Framework. In: Gianola, A., Guizzardi, R., Borbinha, J., Mira da Silva, M., Barateiro, J. (eds) Enterprise Design, Operations, and Computing. EDOC 2025. Lecture Notes in Computer Science, vol 16213. Springer, Cham. https://doi.org/10.1007/978-3-032-15140-7_16

1. Introduction

Face à la profusion d'ontologies de cybersécurité, une comparaison structurée est cruciale pour guider la sélection d'une ontologie pour un cas d'usage donné. Le cadre Framework for Ontology Classification (F4OC) (Martins et al., 2022) offre une approche complète en ce sens. Pourtant, un paradoxe subsiste : bien qu'essentielles à l'interopérabilité des plateformes pour l'orchestration, l'automatisation et la réponse de sécurité (SOAR) et à la gestion des menaces, ces ontologies sont peu réutilisées en pratique. Ce manque d'adoption ne relève pas seulement de lacunes techniques, mais d'un déficit de crédibilité que les métriques d'évaluation classiques ignorent. Nous abordons ainsi deux questions : (1) pourquoi cette faible réutilisation et (2) comment définir et évaluer la crédibilité d'une ontologie ?

2. État de l'Art et Analyse via le Cadre F4OC

Pour répondre à la première question, nous avons étendu une revue de la littérature en couvrant la période post-2021, identifiant 52 nouvelles ontologies en plus des 35 déjà cataloguées par (Martins et al., 2022). Ces ontologies ont été passées au crible du cadre F4OC, qui évalue les ontologies selon trois axes orthogonaux :a) Niveau d'application : Distinction entre ontologies de référence (pour l'échange de connaissances) et opérationnelles (pour l'implémentation) (Guizzardi, 2007);b) Niveau de généralité : Catégorisation en ontologies fondationnelles, de domaine, de tâche ou d'application (Guarino, 1994 ; Van Heijst et al., 1997) ;c) Ancrage fondationnel : Selon que l'ontologie hérite ou non de concepts d'une ontologie fondationnelle (comme UFO ou DOLCE (Guizzardi, 2005 ; Oltramari etal., 2014)).L'analyse révèle un manque critique d'ancrage fondationnel : seules 4 % desontologies sont ancrées alors que 87 % sont opérationnelles. Hormis des exceptions rigoureuses comme COVER (Sales et al., 2018) et ROSE (Oliveira et al., 2022), la majorité (incluant MITRE D3FEND, (Kaloroumakis et Smith, 2021)) manque d'assise conceptuelle, générant des ambiguïtés sémantiques. Enfin, l'hyperspécialisation des ontologies pour des cas d'usage précis constitue un autre frein majeur à leur réutilisabilité.

3. Concept de crédibilité : une nouvelle dimension d'évaluation

Pour dépasser les limites techniques, l'article formalise la crédibilité (ISO/IEC25012) comme le degré de confiance communautaire. Jugeant les métriques existantes trop subjectives, nous proposons quatre dimensions tangibles liées aux parties prenantes : soutien institutionnel, reconnaissance académique, validation des praticiens et adoption industrielle. Cette approche distingue les ontologies théoriques (COVER, ROSE) des modèles pragmatiques (WAVED, UCO) qui s'imposent par leur alignement sur les standards et leur large adoption.

4. Applications du cadre : le cas du projet ANCILE

L'application de notre cadre F4OC étendu avec le critère de crédibilité au projet ANCILE (SOAR autonomique) montre la pertinence de cette approche. Le filtrage technique strict ne retenant que l'ontologie CRATELO, qui n'est plus disponible donc inutilisable, nous avons relaxé les critères d'ancrage. L'intégration de la crédibilité a alors privilégié WAVED : bien que théoriquement imparfaite, son alignement industriel (ATT&CK, D3FEND) la rend opérationnellement intéressante comparée aux candidats purement académiques.

5. Conclusions et perspectives

En enrichissant le cadre F4OC par l'intégration de critères de crédibilitémultidisciplinaires, cette approche offre aux ingénieurs et chercheurs un

outil de sélection d'ontologies de cybersécurité plus fiable et mieux adapté à leurs exigences pratiques.

Remerciements : Ce travail a été financé par l'Agence nationale de la recherche (ANR) dans le cadre du projet ANCILE (subvention n° ANR-23-CE39-0010).

Bibliographie non numérotée

Akbar K., Rahman F., Singhal A., Khan L., Thuraisingham B. (2023). The design and application of a unified ontology for cyber security. Information Systems Security (ICISS 2023), Springer, Raipur, p. 23-41.

Guarino N. (1994). The ontological level. Philosophy and the Cognitive Sciences, Hölder-Pichler-Tempsky, Vienna.

Guizzardi G. (2007). On Ontology, ontologies, Conceptualizations, Modeling Languages, and (Meta)Models. Databases and Information Systems IV, IOS Press, Amsterdam, p. 18-39.

Guizzardi G. (2005). Ontological Foundations for Structural Conceptual Models. Thèse de doctorat, University of Twente.

Kaloroumakis P. E., Smith M. J. (2021). Toward a Knowledge Graph of Cybersecurity Countermeasures. Rapport technique, The MITRE Corporation.

Martins B. F., Serrano L. J., Reyes Román J. F., Panach J. I., Pastor O., Hadad M., Rochwerger B. (2022). A framework for conceptual characterization of ontologies and its application in the cybersecurity domain ; Software and Systems Modeling, vol. 22, n° 1, p. 1-27.

Oliveira Í., Sales T. P., Baratella R., Fumagalli M., Guizzardi G. (2022). An Ontology of Security from a Risk Treatment Perspective. Conceptual Modeling (ER 2022), Springer, Hyderabad, p. 279-293.

Oltramari A., Cranor L. F., Walls R. J., McDaniel P. (2014). Building an ontology of cyber security. Proceedings of the 9th International Conference on Semantic Technologies for Intelligence, Defense, and Security (STIDS 2014), CEUR-WS, Fairfax, p. 54-61.

Sales T. P., Baião F., Guizzardi G., Almeida J. P. A., Guarino N., Mylopoulos J. (2018). The Common Ontology of Value and Risk. Conceptual Modeling (ER 2018), Springer, Xi'an, p. 121-135.

Syed Z., Padia A., Finin T., Joshi A., Peng Y. (2016). UCO: A Unified Cybersecurity Ontology. Proceedings of the AAAI Workshop on Artificial Intelligence for Cyber Security, AAAI Press, Phoenix.

Van Heijst G., Schreiber A. T., Wielinga B. J. (1997). Using Explicit Ontologies in KBS Development ; International Journal of Human-Computer Studies, vol. 46, n° 2-3, p. 183-292.

Wilson S. I., Navaratne C. M., Dharmaratne A. T., Gregory M. A. (2022). Towards a usable ontology: Quality characteristics for an ontology-driven DSS ; IEEE Access, vol. 10, p. 11956-11977.

Pourquoi ouvrir un modèle d'IA générative ? Une typologie fondée sur ce qui est ouvert et ce qui ne l'est pas.

Robert Viseur[1], Nicolas Jullien[2]

1. Service TIC @ FWEG, UMONS
17 Place Warocqué, B-7000 Mons, Belgique
robert.viseur@umons.ac.be

2. Labo LEGO, IMT Atlantique
Technopole Brest Iroise F-29238 Brest, France
nicolas.jullien@imt-atlantique.fr

RÉFÉRENCE DE L'ARTICLE RCIS. Ce texte est un résumé de l'article : Why open a generative AI model? A typology based on what is open and what is not. Cette recherche analyse 189 modèles d'IA générative afin de dépasser l'opposition entre open-weight et open-source. Elle met en évidence trois dimensions de l'ouverture (reproductibilité, readiness, productization) et cinq profils distincts (open washing, easy access, open weight, open science, open source).

MOTS-CLÉS : GenAI, Openness, Open Weight, Open Source, Clustering.

La montée en puissance de l'IA générative a remis au premier plan la question de l'ouverture des modèles. Pourtant, les termes mobilisés pour qualifier cette ouverture demeurent ambigus. En pratique, l'expression « *open source AI* » désigne souvent des réalités hétérogènes. Certains modèles ne rendent publics que leurs poids, d'autres diffusent aussi du code, de la documentation ou des informations sur les données d'entraînement. L'article part de ce constat pour montrer que l'opposition usuelle entre modèles « *open weight* » et « *open source* » ne suffit pas à décrire la diversité empirique des stratégies d'ouverture. La question de recherche est la suivante :

Quelles dimensions structurent l'ouverture des modèles d'IA générative, et quelle typologie de profils d'ouverture émerge au-delà de la distinction entre open-weight et open-source ?

Pour y répondre, nous nous appuyons sur la grille de Liesenfeld *et al.* (2023), qui évalue l'ouverture selon 14 critères répartis en trois catégories : « *availability* » (sources de données, poids, code d'entraînement), « *documentation* » (documentation du code et du matériel, prépublication, article, *model card*, *datasheet*) et « *access methods* » (*package*, API, licences). Le jeu de données

provient de l'Open Source AI Index[1] et porte sur 189 modèles. Les données ont été reconstruites à partir de fiches YAML extraites automatiquement, puis transformées en variables ordinales codées sur une échelle 0/1/2 correspondant à « *closed* », « *partial* » et « *open* ». Cette préparation permet de passer d'un ensemble descriptif hétérogène à une matrice exploitable statistiquement. La méthodologie repose sur l'algorithme HCPC (*Hierarchical Clustering on Principal Components*), implémenté dans R avec FactoMineR. Cette méthode combine réduction de dimension et classification. Une ACP est d'abord réalisée sur les variables ordinales traitées comme numériques ordonnées. Une classification hiérarchique ascendante selon Ward est ensuite appliquée sur les composantes principales. Enfin, une consolidation par « *k-means* » affine la partition. Le HCPC est ici mobilisé pour produire une typologie empiriquement fondée des formes d'ouverture.

Notre recherche permet la mise en évidence de trois dimensions. La première, expliquant environ 27 % de la variance, est la « reproductibilité ». Elle renvoie à la disponibilité des éléments permettant l'audit et la reconstruction du modèle, notamment le code, les sources de données, les *datasheets* et la documentation matérielle. La deuxième, expliquant environ 11 %, est la « *readiness* ». Elle correspond à la disponibilité du modèle pour l'usage, portée par les poids, les licences, l'API et certains signaux de publication. La troisième, expliquant moins de 10 %, est la « *productization* ». Elle distingue les modèles diffusés sous forme d'artefacts téléchargeables de ceux orientés vers des canaux d'intégration comme l'API ou le *packaging*. À partir de ces trois dimensions, nous distinguons cinq grands profils d'ouverture. Le premier cluster, « *open washing* », rassemble des modèles qui donnent les signes extérieurs de l'ouverture sans en assumer pleinement les exigences. L'ouverture y est davantage affichée que réellement mise en œuvre. Le deuxième, « *easy access* », privilégie avant tout la facilité d'accès et d'usage, grâce aux API ou aux *packages*, sans pour autant s'accompagner d'une transparence équivalente sur les conditions de production du modèle. Le troisième, « *open weight* », correspond à une ouverture centrée sur la mise à disposition des poids, mais qui laisse en retrait les éléments nécessaires à une véritable reproductibilité. Le quatrième, « *open science* », s'inscrit dans une logique plus académique, marquée par l'importance accordée à la documentation, au code et aux artefacts de recherche. Enfin, le cinquième cluster, « *open source* », représente la forme la plus aboutie d'ouverture, en combinant diffusion des poids, accès au code, informations sur les données et documentation suffisamment riche pour soutenir l'audit, la réutilisation et la reconstruction du modèle.

La partition en cinq classes est retenue comme le meilleur compromis entre qualité statistique et interprétation. Les frontières entre *clusters* ne sont pas totalement nettes, mais ils définissent des profils suffisamment stables pour être pertinents sur le plan analytique. L'apport théorique de l'article est de montrer que l'ouverture des modèles génératifs ne se réduit ni à une opposition entre fermé et ouvert, ni à la seule distinction entre « *open weight* » et « *open source* ». Elle est en effet motivée par des objectifs tels que l'adoption, l'auditabilité, l'exploration ou la reconstruction.

[1] Cf. https://osai-index.eu/.

Impact environnemental des processus métier, analyse des impacts embarqués et d'usage des ressources

Matteo Ciccone[1], Mario Cortes-Cornax[1], Agnès Front[1], Claudia Roncancio[1]

1. Laboratoire d'Informatique de Grenoble
Université Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG
Grenoble, France Labo, Université AR_adelec1
prenom.nom@univ-grenoble-alpes.fr

REFERENCE DE L'ARTICLE Ce texte est un résumé de l'article : Analyzing Embodied and Use-Phase Environmental Impacts of Resources within Business Processes – Matteo Ciccone, Mario Cortes-Cornax, Agnès Front, Claudia Roncancio – 20th International Conference on Research Challenges in Information Science – RCIS 2026 – Toulouse, France – May, 2026.

MOTS-CLES : Processus métier, Impact environnemental, Analyse de Cycle de Vie, Ressources

La crise environnementale appelle à une analyse critique des impacts environnementaux des activités humaines à l'échelle mondiale. Notre travail se concentre sur les organisations cherchant à évaluer et améliorer l'impact environnemental de leurs processus métier (PM). Il contribue aux efforts de recherche du *Green BPM* (Couckuyt et Van Looy, 2020) (Gohar et al, 2020) qui étendent les méthodes et outils BPM dans le but d'intégrer les indicateurs environnementaux et les considérations environnementales à chaque étape de la gestion des processus.

Dans des travaux précédents (Cortes-Cornax et al., 2025), nous nous sommes concentrés sur la gouvernance des systèmes d'information en proposant la méthode *GreenPath*, dont l'objectif principal est l'analyse des impacts environnementaux des processus métiers des organisations dans un but de réduction. *GreenPath* adopte une approche systémique qui couvre plusieurs niveaux d'analyse à savoir l'organisation, les processus, les activités et les ressources. Elle permet la création d'une vue environnementale, appelée E-view, du modèle d'un processus métier, en clarifiant les ressources utilisées et leur information environnementale. Nous suivons une approche d'analyse du cycle de vie (ACV) (ISO, 2006).

En alignant l'analyse des ressources avec la logique métier, il est possible de mieux identifier des leviers pour réduire les impacts environnementaux négatifs. Pour y parvenir, il est nécessaire de prendre en compte l'empreinte environnementale

générée pendant toutes les phases du cycle de vie des ressources, en incluant leur fabrication, leur transport, leur utilisation et potentiellement leur destruction en fin de vie. Lors de l'analyse des ressources, en particulier des ressources informatiques, les travaux connexes se concentrent principalement sur les impacts associés à la phase d'usage, en termes de consommation d'énergie. Néanmoins, négliger l'impact des autres phases mentionnées ci-dessus, appelé *impact embarqué*[1], peut conduire à une sous-estimation significative de l'impact environnemental réel des processus métier. Par exemple, Boavizta[2] estime que la fabrication d'un serveur informatique produit en moyenne 800 kgCO2eq, alors que d'après (Hirz et Nguyen, 2002), la fabrication d'une voiture électrique produit en moyenne 14000 kgCO2eq. Ignorer ces impacts conduirait à des estimations inexactes.

La principale contribution de cet article est la formalisation de stratégies d'allocation de l'impact embarqué des ressources aux différents niveaux d'une organisation. Ces stratégies définissent des règles pour la répartition des impacts intrinsèques des ressources entre activités ou processus. Trois types de stratégies sont formalisées. ***Equal split*** propose une répartition égale entre toutes les activités utilisant la ressource. ***Usage based*** effectue une répartition proportionnelle à l'usage fait en termes de temps ou de service rendu. ***Life-Span based*** définit une répartition prenant en compte la durée de vie attendue des ressources.

Le choix de la stratégie dépend des caractéristiques des ressources, des conditions contextuelles d'utilisation et des objectifs de l'analyse. La formalisation de ces stratégies est essentielle pour une meilleure compréhension des impacts environnementaux réels des processus métier et au final, pour améliorer leurs impacts.

Bibliographie

Cortes-Cornax, M., Front, A., Oliveira, R., Roncancio, C. (2025). A method for assessing environmental impacts of business processes. In: 27th International Conference on Business Informatics (CBI'25).

Couckuyt, D., Van Looy, A. (2020). A systematic review of green business process management. Business Process Management Journal 26(2), 421–446.

Gohar, S.R., Indulska, M., et al. (2020). Environmental sustainability through green busi-ness process management. Australasian Journal of Information Systems 24.

Hirz, M., Nguyen, T.T. (2022). Life-cycle co2-equivalent emissions of cars driven by conventional and electric propulsion systems. World electric vehicle journal 13(4), 61.

International Organization for Standardization (2006). Iso 14040:2006 – environmentalmanagement – life cycle assessment – principles and framework.

[1] *embodied impact* en anglais
[2] https://datavizta.boavizta.org/serversimpact

De l'Archive au Graphe de Connaissances : Post-correction OCR et Extraction d'Entités par LLMs sur les Fonds Historiques du Cnum

Mohamed Amine Lasheb, Olivier Pons

Laboratoire CEDRIC, Conservatoire national des arts et métiers
292 Rue Saint-Martin, 75003 Paris, France
mohamed-amine.lasheb@lecnam.net, olivier.pons@lecnam.net

RESUME. *La valorisation des fonds historiques se heurte à la qualité variable de l'OCR. Cet article compare des modèles locaux et commerciaux sur trois siècles ($XVIII^{e}$-XX^{e}). Si Qwen2.5-72B est performant sur le XXe siècle (51,8% de correction), Gemini-2.5-Flash s'impose sur les documents du XVIIIe (62,7% de correction) en gérant mieux les spécificités typographiques anciennes. Nos résultats démontrent un effondrement sémantique des modèles locaux sur de longs contextes, amorcé dès 2k tokens et devenant critique au-delà de 4k tokens.*

ABSTRACT. *The valorization of historical document collections is hindered by variable OCR quality. This paper compares local and commercial large language models across three centuries (18^{th}–20^{th}). While Qwen2.5-72B performs strongly on 20th-century documents (51.8% correction rate), Gemini-2.5-Flash outperforms on 18th-century materials (62.7% correction rate), demonstrating superior handling of archaic typography such as the long s. Our results reveal a semantic collapse in local models when processing long contexts, emerging as early as 2k tokens and becoming critical beyond 4k tokens.*

MOTS-CLÉS : *OCR, LLM, Humanités Numériques, Graphes de Connaissances, Extraction d'Information, Ingénierie Documentaire*

KEYWORDS: *OCR, LLM, Digital Humanities, Knowledge Graphs, Information Extraction, Document Engineering*

1. Introduction

Le Conservatoire Numérique des Arts et Métiers (Cnum)[1], pionnier des bibliothèques numériques (Hohnsbein *et al.*, 2025), conserve une collection majeure documentant l'histoire des techniques du XV^e siècle à nos jours. Nous nous concentrons ici sur la période du $XVIII^e$ au XX^e siècle, charnière marquée par l'évolution des glyphes et l'industrialisation de l'imprimerie.

Face à ce corpus massif, l'ingénierie documentaire doit s'automatiser. Le catalogue maître (corpus *chapeau*[2]) recense **2 447 monographies**, réparties en **3 029 unités numériques** (corpus *QUANTA*[3]), totalisant des dizaines de milliers de pages brutes accessibles via le protocole standard d'interopérabilité, OAI-PMH.

Cependant, la qualité variable de l'OCR freine cette valorisation. Les spécificités typographiques anciennes génèrent un bruit numérique important (ex : « s long » lu « f ») qui complique l'indexation sémantique et rend inopérantes les chaînes classiques des SI orientés graphes, ingérant alors des nœuds erronés.

Pour y remédier, notre démarche s'articule avec le projet **ANR LAURA**[4]. Nous fournissons une brique technologique en amont : un pipeline LLM de nettoyage et d'extraction capable d'alimenter ces bases en « factoïdes » propres (assertions historiques structurées en n-uplets, ex : [Sujet - Relation - Objet - Date]).

2. État de l'art

Extraire des informations de textes historiques dits « bruités » reste un défi documenté. Ce bruit résulte de la dégradation des supports et de l'incapacité des moteurs OCR à interpréter les typographies anciennes. L'utilisation de pipelines TAL classiques pour peupler une base de connaissances affiche des performances très faibles (22% de précision pour les personnes), un échec directement imputable à ce bruit OCR (Quaresma and Finatto, 2020). De plus, si l'exploration sémantique par vectorisation et RAG (*Retrieval-Augmented Generation*) a récemment démontré son potentiel pour interroger de vastes corpus anciens, elle reste tributaire de la qualité de la segmentation et de l'indexation initiale (Blangeois *et al.*, 2024).

Pour pallier ce problème en amont, avant l'avènement des modèles génératifs, des approches de traduction *sequence-to-sequence* (LSTM) en deux étapes ont montré de bons résultats (Schaefer and Neudecker, 2020). Aujourd'hui, la post-correction par Grands Modèles de Langage (LLMs) est explorée comme alternative (Boros *et al.*, 2024). Toutefois, un risque élevé d'hyper-correction subsiste : le modèle tend à réécrire le texte en syntaxe moderne, détruisant la source historique. Notre étude évalue la capacité des LLMs à assainir l'OCR tout en préservant l'archive (c'est-à-dire sans

1. https://cnum.cnam.fr
2. Notice bibliographique décrivant l'œuvre dans sa globalité.
3. Unité numérisée (tome, fascicule, planche, etc.) rattachée à une notice chapeau.
4. https://anr.fr/Projet-ANR-25-CE38-0700

moderniser la syntaxe d'époque), afin de définir une architecture SI optimale où un seul modèle conjoint nettoie puis extrait directement les entités.

3. Méthodologie

Notre pipeline (Figure 1) agit en trois étapes séquentielles : 1) le moissonnage des textes OCRisés via OAI-PMH, 2) la segmentation en blocs de 2k à 4k tokens pour une post-correction générative (lissant le bruit tout en préservant la fidélité historique), et 3) l'extraction structurée.

Pour l'extraction, nous écartons les modèles NER classiques (limités à la détection de termes isolés)[5] qui exigeraient un système complexe d'extraction de relations (RE) supplémentaire. Nous privilégions une approche conjointe par LLM avec un prompt *Zero-Shot* : le modèle est instruit d'agir comme un historien expert pour générer directement un graphe de factoïdes au format JSON (sujet, relation, objet, date)[6].

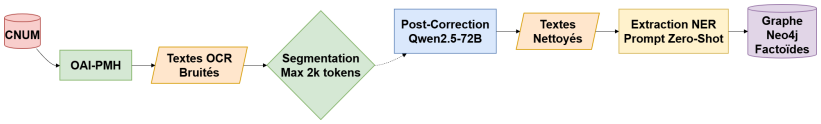


FIGURE 1. *Architecture du SI : post-correction OCR et extraction de factoïdes.*

Pour évaluer nos modèles face à l'hétérogénéité des fonds, nous avons constitué un corpus de test de 120 pages, dont l'OCR initial a été généré par **ABBYY FineReader Server 14**. Ces pages ont été échantillonnées aléatoirement sur trois époques : 30 pages de *L'art du menuisier ébéniste* (1774), 30 pages du *Portefeuille de l'ingénieur des chemins de fer* (1843), et 60 pages de *La Science et la Vie* (1913-1945). La constitution de la Vérité Terrain (transcription et correction manuelles) a nécessité environ **deux jours d'effort d'annotation**.

Pour les documents des $XVIII^{e}$ et XIX^{e} siècles, caractérisés par un bruit OCR dense et des glyphes obsolètes tels que le *s long*, l'évaluation a été réalisée page par page. Cette approche évite la perte de contexte observée lors du traitement par lots (*batching*) sur des textes hautement dégradés.

Afin de comparer les approches souveraines (locales) et commerciales (Cloud), nous avons évalué cinq modèles. Les inférences locales ont été réalisées sur un serveur équipé de deux cartes graphiques NVIDIA RTX 6000 Ada Generation (48 Go de VRAM chacune). Parmi les solutions testées figurent **Qwen2.5-72B** et **32B**, des modèles locaux denses reconnus pour leur multilinguisme, ainsi que **DeepSeek-R1-70B**, un modèle local orienté vers le raisonnement. Nous avons également intégré **GPT-OSS 120B**[7], un

5. Ex : modèle `Jean-Baptiste/camembert-ner-with-dates` ou GLiNER.
6. Code, corpus et prompts détaillés sur : `https://github.com/aminelasheb/Arch2KG`
7. `https://ollama.com/library/gpt-oss`

modèle local de très grande taille. Enfin, pour représenter les solutions Cloud, nous avons utilisé **Gemini-2.5-Flash**, un modèle commercial via API. Ce dernier a toutefois été exclu des mesures de temps d'inférence en raison de la variabilité inhérente aux latences réseau et aux serveurs tiers.

L'évaluation de la post-correction OCR pose un défi méthodologique : comparer directement la prédiction du LLM à la vérité terrain génère des scores artificiellement hauts ($> 90\%$), car la majorité du texte source est déjà correcte. Pour isoler l'impact réel du modèle, nous avons implémenté un **alignement à trois voies (3-way alignment)** au niveau du mot. Nous identifions d'abord les erreurs exactes de l'OCR (E_{ocr}) et les mots corrects (C_{ocr}) par rapport à la vérité terrain. Nous calculons ensuite deux métriques strictes : le **Taux de Correction** (Vrais Positifs), qui mesure la proportion des erreurs E_{ocr} que le LLM a réussi à corriger parfaitement, et le **Taux de Dégradation** (Faux Positifs / Hallucinations), qui quantifie la proportion des mots initialement corrects C_{ocr} que le LLM a altérés ou réécrits à tort.

L'évaluation de l'extraction d'entités (NER) sur des documents historiques ne peut se satisfaire d'une métrique stricte (*Exact Match*). Le bruit résiduel pénalise artificiellement les modèles. Nous avons implémenté une évaluation par **Correspondance Partielle** : une entité prédite est considérée comme un Vrai Positif si elle est une sous-chaîne exacte de la vérité terrain, ou inversement (ex : "12 fr." validé par "Abonnement 12 fr."). Cette tolérance reflète la viabilité réelle des données pour une base de connaissances.

4. Résultats et Discussion

L'évaluation (Figure 2) montre que sur le corpus moderne (XX^{e}), **Qwen2.5-72B** rivalise avec le cloud avec **51,8%** de correction sur des segments de 2k tokens, confirmant la viabilité des modèles souverains pour la typographie standardisée.

Cependant, sur les fonds anciens ($XVIII^{e}$-XIX^{e}), les modèles locaux plafonnent (**49,7%** pour DeepSeek, **45,8%** pour Qwen) face au modèle **Gemini-2.5-Flash** (**62,7%**). Ce décalage s'explique par la mauvaise interprétation du *s long*, souvent confondu avec « *f* », « *I'* » ou « *T* » (ex : *favoir* au lieu de *savoir*).

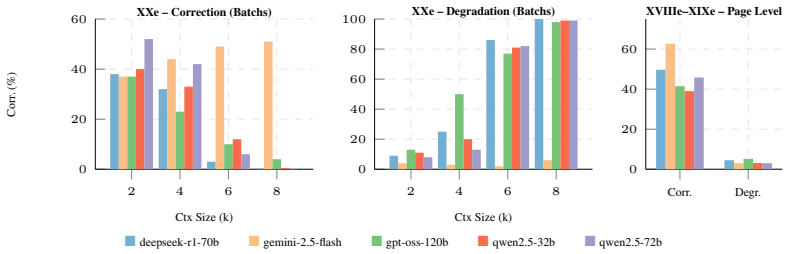


FIGURE 2. *Performances OCR : comparaison séculaire et impact de la fenêtre de contexte.*

Enfin, un « mur du contexte » apparaît au-delà de 4k tokens : les modèles locaux subissent alors un effondrement sémantique (hallucinations massives), tandis que l'API Gemini maintient sa stabilité jusqu'à 8k tokens, rendant la segmentation courte impérative pour les architectures locales.

La Figure 3 illustre les scores F1, de Précision et de Rappel des modèles locaux en contrastant les résultats obtenus sur les batchs de 2k et 4k tokens (les fenêtres plus larges ayant été écartées au vu de la dégradation OCR).

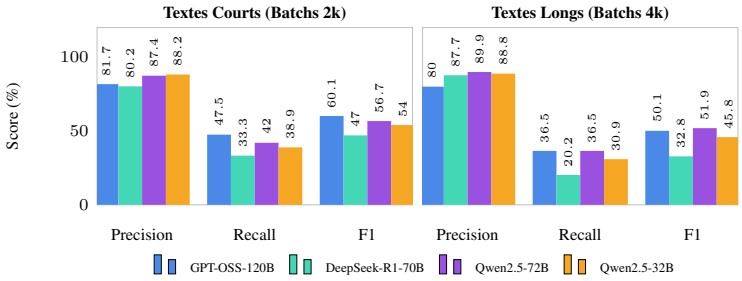


FIGURE 3. *Performances NER des modèles locaux (Correspondance Partielle).*

L'analyse NER révèle une forte asymétrie. La précision est excellente (>80%) : les modèles locaux n'hallucinent presque pas, garantissant des factoïdes fiables. Le rappel, en revanche, dépasse rarement 45%. Face à de longues énumérations, les LLMs souffrent d'une « paresse » générative et tronquent leurs extractions, un défaut exacerbé sur les contextes de 4k tokens. Au global, **Qwen2.5-72B** et **GPT-OSS-120B** s'imposent comme les solutions locales les plus équilibrées (F1 $\approx$ 56-60% sur 2k tokens), offrant le meilleur compromis pour automatiser la collecte dans le cadre du projet LAURA. À titre d'illustration, notre pipeline génère directement des factoïdes structurés prêts à être injectés dans un SI orienté graphe. Pour une phrase comme « Louis Pasteur invente le vaccin en 1885 », le modèle produit : `{"sujet": "Louis Pasteur", "relation": "invente", "objet": "vaccin", "date": "1885"}`, assurant une interopérabilité immédiate sans surcouche d'extraction de relations.

L'industrialisation du pipeline exigeant d'évaluer sa faisabilité temporelle, le TABLEAU 1 détaille les temps d'inférence sur notre architecture bi-GPU.

TABLEAU 1 – TEMPS D'INFÉRENCE MOYEN PAR BLOC (SEC) SUR UN BI-GPU RTX 6000 ADA.

	Post-correction OCR		**Extraction NER**	
Modèle Local	**Batch 2k**	**Batch 4k**	**Batch 2k**	**Batch 4k**
Qwen2.5-32B	41.1 s	92.7 s	14.8 s	28.5 s
Qwen2.5-72B	87.3 s	210.4 s	36.2 s	72.5 s
DeepSeek-R1-70B	114.9 s	215.5 s	67.1 s	101.4 s
GPT-OSS-120B	33.8 s	59.6 s	39.9 s	53.9 s

Ces temps valident un traitement asynchrone local : **Qwen2.5-72B** (87,3 s) est le modèle optimal et **Qwen2.5-32B** le meilleur compromis (les hallucinations disqualifiant GPT-OSS). Par ailleurs, l'industrialisation soulève des enjeux de gouvernance SI. Pour l'éthique patrimoniale, le SI doit conserver l'archive brute et versionner la correction LLM. Les modèles souverains y répondent parfaitement : ils garantissent la confidentialité des fonds, l'indépendance face au cloud et la durabilité de l'infrastructure.

5. Conclusion et Perspectives

L'intégration de l'IA générative dans les SI patrimoniaux impose des choix d'architecture stricts. Nos travaux ($XVIII^{e}$-XX^{e}) démontrent que pour le fonds moderne, le modèle local **Qwen2.5-72B** s'impose comme la solution souveraine de référence sur des segments de 2k tokens. Pour les fonds anciens, bien que les modèles locaux soient acceptables, ils butent sur la typographie de l'Ancien Régime, là où le modèle commercial **Gemini-2.5-Flash** maintient une haute robustesse.

L'analyse du « mur du contexte » prouve qu'au-delà de 2k tokens, les modèles locaux amorcent une instabilité qui devient critique à 4k tokens, rendant la segmentation courte indispensable. Pour les archives les plus anciennes, les performances actuelles des modèles locaux constituent une base prometteuse qui pourra être surpassée par une approche **multimodale native**. Le couplage futur du texte avec l'image source (vision) permettra de capter avec précision les spécificités graphiques aujourd'hui mal interprétées. Ce pipeline constitue la première brique d'un projet de graphe de connaissances pour les humanités numériques, soutenu par le personnel du Cnum que nous remercions pour l'accès facilité à ces fonds historiques.

Bibliographie

Blangeois M. *et al.*, « L'exploration de l'Encyclopédie par l'intelligence artificielle : problèmes, méthodes, premiers résultats », *Qu'est-ce que l'I.A peut faire pour vous ?*, 2024.

Boros E. *et al.*, « Post-Correction of Historical Text Transcripts with Large Language Models : An Exploratory Study », *Proc. of LaTeCH-CLfL 2024*, 2024.

Hohnsbein A. *et al.*, « Aux origines du Conservatoire numérique des arts et métiers (Cnum). Entretien avec Pierre Cubaud et Cécile Formaglio », *Cahiers d'histoire du Cnam*, 2025.

Quaresma P., Finatto M. J. B., « Information Extraction from Historical Texts : a Case Study », *Proc. of DHandNLP 2020*, 2020.

Schaefer R., Neudecker C., « A two-step approach for automatic OCR post-correction », *Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, p. 52-57, 2020.

Phroneo : A Domain-Agnostic Agentic GraphRAG Architecture for Multi-Hop Questions

Gloria Elena Jaramillo Rojas[1], Meriem Sabrine Halilali[1], Rialy Andriamiseza[1]

1. Capgemini Engineering
11 Av. Didier Daurat, 31700 Blagnac, France
gloria-elena.jaramillo-rojas@capgemini.com, meriem.halilali@capgemini.com, rialy.andriamiseza@capgemini.com

RESUME. Les organisations gèrent un volume d'informations de plus en plus important ; toutefois, la richesse de ces informations dépend principalement de leur capacité à être efficacement accessibles et exploitables. Les systèmes traditionnels de recherche par mots-clés et de questions-réponses à une seule étape (single-hop) restent insuffisants pour les requêtes complexes nécessitant un raisonnement sur plusieurs documents. Cet article présente Phroneo, une solution graphRAG multi-agent conçue pour répondre à un besoin industriel identifié : la recherche documentaire à partir de documents techniques. Phroneo intègre deux composants principaux : un module d'extraction et de représentation automatiques des connaissances, et un module d'orchestration d'agents pour la décomposition des questions, le raisonnement et la génération de réponses. L'intégration d'une stratégie graphRAG basée sur un graphe de connaissances enrichi permet des réponses vérifiables et contextualisées, réduisant les réponses erronées et améliorant la fiabilité. De plus, sa conception modulaire, fondée sur les fonctions cognitives des agents, facilite l'extension à différents types de questions techniques multi-étapes (multi-hop).

ABSTRACT. Organizations manage an increasingly large volume of information; however, the richness of this information depends mainly on its ability to be effectively accessed and exploited. Traditional keyword-based search and single-hop question-answering systems remain insufficient for complex queries that require reasoning across multiple documents. This article presents Phroneo, a multi-agent graphRAG solution designed to address an identified industrial need: document search using technical documents. Phroneo integrates two main components: a module for automatic knowledge extraction and representation, and an agentic orchestration module for question decomposition, reasoning and response generation. The integration of a graphRAG strategy based on an enriched knowledge graph enables verifiable and context-aware responses, reducing hallucinations and improving reliability. Moreover, its modular design based on agent's cognitive functions allows for an easy extension of different types of multi-hop technical questions.

MOTS-CLES : questions multi-étapes, systèmes agentiques, LLM, graphe de connaissances, ontologie

KEYWORDS: multi-hop questions, agentic systems, LLM, knowledge graph, ontology

1. Introduction

Query-answering (QA) systems have become a cornerstone of modern information retrieval, enabling users to access relevant knowledge efficiently across vast datasets. Their success lies in the ability to interpret user queries and extract precise answers from structured or unstructured sources. However, single-hop questions-answering systems lack of efficient mechanisms to support complex information retrieval needs.

Multi-hop questions represent a class of queries that require reasoning across multiple pieces of information, often distributed across different documents or knowledge bases. Unlike single-hop questions, which can be answered by retrieving a single fact, multi-hop questions demand a chain of reasoning steps and link facts coherently. This dependency on intermediate reasoning introduces challenges such as understanding the semantic of the question, maintaining context and handling ambiguity.

Traditional question-answering systems struggle with these requirements because they are optimized for direct fact retrieval rather than iterative reasoning. Their architectures often lack mechanisms for decomposing complex queries, coordinating multiple reasoning steps or they are based on strict assumptions such as the sequentiality of the sub-questions. As a result, they fail to deliver accurate answers when information is fragmented. The need for robust multi-hop question-answering systems reflects a growing demand in industry. Architectures able to handle these complex queries are a priority for both researchers and industry (Abir, 2024; Yuntao, *et al.*, 2022; Zhang, *et al.*, 2024).

This paper presents a multi-agent system called Phroneo specially designed for answering multi-hop questions. Phroneo has been developed to address a specific industrial need of one of our clients: enabling advanced documentary retrieval across multiple sources. Existing commercial systems remain insufficient for such a need, as they do not adequately satisfy the client's operational constraints, particularly with respect to the integration and exploitation of fragmented knowledge distributed across several documents. The article is structured as follows: section 2 presents the literature review. The proposed architecture and components are described in Section 3. The discussion and conclusions are presented in Sections 4 and 5, respectively.

2. Literature Review

In (Ghafouri, *et al*., 2025), authors highlight the challenges of Knowledge Graph Question Answering for Persian. In this work, authors propose an architecture for answering multi-hop questions based on KG. The architecture is composed of 4 components: the question decomposer, the entity recognition and linking components, the SPARQL query generator, and the query executor. Although questions are firstly decomposed into simpler segments, authors based their work on the sequential linking of identified entities in the knowledge graph (KG), which restricts the type of complex questions handled by the architecture. Moreover, they restrict the extraction of the

knowledge to an exact match by the query. On the other hand, the sequential reasoning of this work is overcome in (Bian, *et al.,* 2025) by proposing a LLM to tackle the non-linearity in the reasoning of multi-hop question answering. Instead of relying on a static KG, the authors propose the creation of a question-centric knowledge graph ontology. This approach is composed of three sequential stages: ontology extraction, FOL construction, and sub-question decomposition. Similarly to our approach, the authors extract not only entities but also the concepts associated with those entities. However, their method does not leverage the agentic capabilities to analyze the sub-questions. Moreover, the answer is linked to a predicted class in the ontology but no information is provided about how to handle more elaborate answers.

GraphTrace (Osipjan, *et al.,* 2025) also proposes the integration of KG with LLM to answer multi-hop questions. Unlike our approach, GraphTrace directly relies on structured data. The architecture behind GraphTrace comprises several modules in charge of entity extraction, path finding, query decomposition, semantic path ranking, context aggregation, and LLM-based answer generation. This work emphasizes the role of multi-hop paths in providing contextual information relevant to the query. However, it relies in the exact matching of the extracted entities in the KG for the path finding.

INRAExplorer (Lelong, *et al.,* 2025) is an agentic graphRAG system for scientific data exploration. The agent can navigate different tools to find different ways to retrieve information. Unlike previous approaches, this work proposes a hybrid knowledge base (vector base and KG). The architecture of INRAExplorer is composed of a single agent in charge of understanding and decomposing the user query, creating an action plan, selecting and invoking the retrieval tool and aggregating the extracted information. No detailed information is provided about how complex questions are decomposed or how the plan is created.

From the literature review, current approaches to answering multi-hop questions based on KG exhibit important limitations, namely: existing systems often depend exclusively on sequential reasoning, exact matching of entities, or predefined ontological structures. Other methods introduce dynamic question-centric ontologies but still require FOL-translatable questions. These gaps highlight the need for an architecture that integrates KG with more flexible reasoning mechanisms.

3. Phroneo's Architecture

Considering our identified industrial need, we propose an agentic graphRAG approach designed to address the challenges of navigating and reasoning over complex technical documentation for which traditional keyword-based search methods and commercial LLM-based solutions are insufficient. The Phroneo's objectives are 1) to reduce the time spent on manual information retrieval 2) to enable verifiable and context-aware responses for critical applications 3) to improve the precision and clarity of answers to complex technical queries.

Our proposed architecture is based on two main modules: 1) Sophion module: for the extraction and representation of knowledge, and 2) An agentic module for the orchestration of specialized agents (Figure 1).

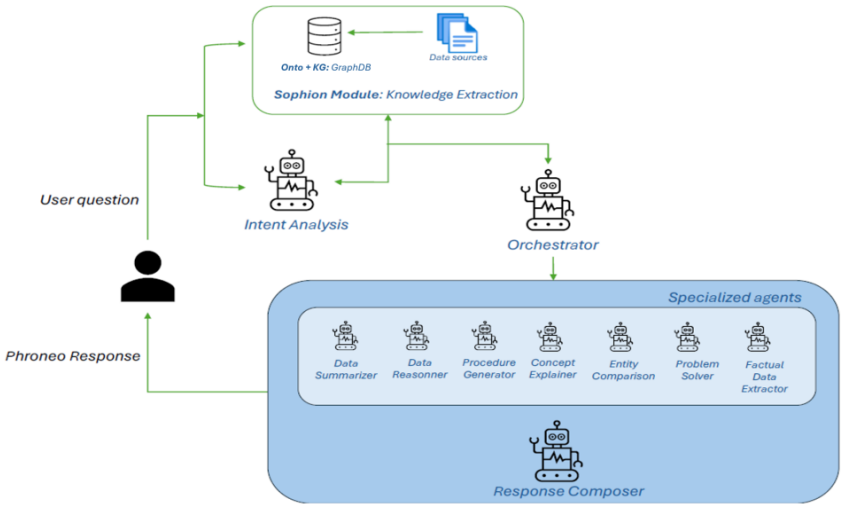


Figure 1. Phroneo's Architecture

3.1. Sophion: Module for the Extraction and Representation of Knowledge

Based on an elaborate prompt given to a LLM, Sophion leverages technical documentation to build an enriched KG based on an automatically generated ontology, which ensures semantic integrity, consistency, and disambiguation of its entities. The ontology is not predefined but emerges progressively from the extraction pipeline, in which candidate entities and relations are first identified in the form of candidate triples from document segments and then normalized and validated through post-processing steps (e.g., correction of malformed triples, normalization of labels, and enforcement of RDF-compliant structures). We intentionally use the term "enriched" graph since we extract metadata for each entity, namely: the entity description, and a list of entity aliases (if available). For the relations in the KG, we extract the relationship description, the relation strength, and the support evidence (quotation extracted from the text justifying the relation), ensuring traceability and interpretability of the generated knowledge graph through reification. We aim our approach to be use-case agnostic. It means we do not need to have previous knowledge about the content of the data source or have a pre-defined ontology.

3.2. Agentic Module: Orchestration of Specialized Agents

This module, implemented using LangChain[1] and LangGraph[2], analyzes the multi-hop question and coordinates the specialized agents via the orchestrator. Initially, the user question is embedded. Then, semantic similarity is calculated on-the-fly to match the user request with the enriched graph to extract the top-10 most relevant nodes. Although, a translation of the user question into queries is also available in Phroneo, the calculation of the semantic similarity allows us to exploit the metadata of the KG. Thus, we propose a solution to the fact of relying on the strict matching of the SPARQL query. We argue that the use of the metadata offers greater semantic richness to the context by capturing nuances more comprehensively. Moreover, to leverage the knowledge representation, a subgraph composed of a neighborhood of region 3 around the relevant node is also extracted. In parallel to this operation, the agentic module is executed. Phroneo is composed of the following agents:

– intention analysis agent: Initially, the user question is analyzed by the "intent analysis" agent, who segments the multi-hop question into sub-logical parts, sometimes reformulating the question. For each segment, the agent is prompted via a LLM to assign an intention (from a pre-defined list) together with the justification of the reasoning.

– orchestrator agent: the orchestrator takes the identified intentions (e.g., reasoning, summarization, comparison, etc.), segments and retrieved knowledge and creates a plan linking the segments and associating them with the agent that is most appropriate for answering each sub-segment (sub-question). For the creation of the plan, the coordinator has the freedom to create parallel, sequential or mixed paths. The list of available agents and capabilities are: (i) data summarizer: this agent is tailored for text summarization tasks; (ii) data reasoner: agent, also used as a fallback mechanism, specifically engineered in solving tasks requiring inferential analysis and logical connections; (iii) procedure generator: this agent focuses on the creation of step-by-step instructions; (iv) concept explainer: agent specialized in the explanation of complex technical ideas; (v) entity comparison agent: agent specialized in the identification and explanation of similarities and differences between concepts; (vi) problem solver: the role of this agent is purpose-built to analyze and resolve issues; and (vii) factual data extractor: retrieves precise information from the context. At the end of the plan execution, the final response composer agent synthesizes the partial outputs into a coherent final response which is sent back to the user.

Unlike existing agentic workflows solutions based on orchestrations, Phroneo does not coordinate agents based on the actions required to answer a user request (route a request towards a LLM or vector store, expand the request, prompt rewriting, or citation generation), but we design agents based on cognitive functions. By cognitive functions, we refer to advanced reasoning capabilities. Each agent is explicitly designed to specialize in one of these functional roles rather than a predefined execution step.

[1] LangChain: https://www.langchain.com/
[2] LangGraph: https://www.langchain.com/langgraph

4. Discussion

The proposed solution combines the strengths of the graph-based retrieval methods with the benefits of agentic systems. It exploits interconnected knowledge through a graphRAG approach, while leveraging agents' autonomy, orchestration and their ability to handle complex problems.

With respect to the Phroneo's objectives introduced in Section 3, adopting a domain agnostic approach provides clear advantages. However, when applied to a completely new domain, our automatic generation of the ontology and the KG requires approximately up to three minutes. Nevertheless, the time required to execute the complete pipeline remains significantly shorter than manual search. The second objective of Phroneo is addressed through the integration of a RAG strategy enriched by metadata from the KG. This design grounds system outputs in verifiable technical documentation, reducing hallucinations and improving reliability. Furthermore, the precision and clarity of the answers are addressed by a better understanding of the semantics of the question. The orchestrator agent builds its execution plan independently of the order or syntax of the user request. This enables agents to better identify intents and delegate each sub request to a specialized agent prompted for handling a specific cognitive function. This modular design and agent specialization facilitate the extension of the system to additional categories of multi-hop queries, including pattern recognition, or statistical analysis, with minimal adjustments.

Despite its advantages, the approach inherits some limitations of LLM-based systems, including computational cost and dependency on external APIs. These aspects must be considered when deploying the system in industrial environments.

In addition, due to the short paper format, we focused on the architectural and conceptual contributions of Phroneo while a quantitative evaluation is currently under development and will include performance metrics such as execution time, quality of extracted entities and relations, and answers relevance.

5. Conclusion and Perspectives

This work introduced Phroneo, an agentic graphRAG architecture designed to address multi-hop technical questions. In our work, we target technical documentation, which is characterized by (i) interdependency in the information: which increases reasoning complexity, (ii) need for reliable answers: specially in domains including safety-critical documentation, such as aerospace, (iii) need of integrate details from several sources: correct interpretation of technical documentation often requires a complex understanding.

The Phroneo's architecture relies on an automatically generated ontology and a knowledge graph to represent technical sources. A set of specialized agents analyzes user intent, decomposes complex queries, and retrieves relevant information from the knowledge graph. An orchestrator agent constructs an execution plan and invokes the agents best suited to handle each segment of the question. A key distinction of our approach is the decomposition of tasks according to cognitive functions rather than

sequential process actions. Phroneo also benefits from rich knowledge extraction supported by metadata included in the ontology and the knowledge graph. Phroneo is designed to be domain-agnostic, which removes the need for predefined ontologies. Moreover, the modularity of the specialized agents enables an easy integration of new capabilities for additional user intents.

Phroneo remains under development, and several directions for future work have been identified. First, improvements in knowledge representation will focus on storing and selecting previously created ontologies. A dedicated agent will pre-analyze new datasets to retrieve an existing ontology when possible. Second, a learning module will be developed to analyze outputs, extracted information, and question segmentation to refine agent behavior over time.

Acknowledgments

I would like to thank our project intern Chloé VARSOVIE and Swapna Sri KOPPULA for their technical contribution to the implementation of Phroneo and Sophion. As well as Bilal CHEHAIBOU, Ba Huy TRAN and Xin HUANG for the initial implementation of the architecture.

Bibliography

Abir C. (2024) *Multi-hop Question Answering over Knowledge Graphs using Large Language Models*. https://arxiv.org/abs/2404.19234

Bian H., Qi Y., Yang R., Che Y., Wang J., Xia H. and Zhen R. (2025). *From Query to Logic: Ontology-Driven Multi-Hop Reasoning in LLMs*. https://arxiv.org/abs/2508.01424

Ghafouri A., Firouzmandi M. and Naderi, H. (2025) *A Method for Multi-Hop Question Answering on Persian Knowledge Graph*. https://arxiv.org/abs/2501.16350

Lelong J., Errazine A. and Blangero, A. (2025). *Agentic RAG with Knowledge Graphs for Complex Multi-Hop Reasoning in Real-World Applications.* https://arxiv.org/abs/2507.16507

Osipjan A., Khorashadizadeh H., Kessel A.-L., Groppe S., and Groppe J. (2025). GraphTrace: A Modular Retrieval Framework Combining Knowledge Graphs and Large Language Models for Multi-Hop Question Answering. *Computers*, vol, 14, n° 9, p. 382-403.

Yuntao K., Phuong N. M., Racharak T., Le T., and Minh N. L. (2022). An Effective Method to Answer Multi-hop Questions by Single-hop QA System. *International Conference on Agents and Artificial Intelligence*, vol. *2*, p. 244-253.

Zhang K., Zeng J., Meng F., Wang Y., Sun S., Bai L., Shen H., and Zhou, J. (2024). Tree-of-Reasoning Question Decomposition for Complex Question Answering with Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, n° 17, p. 19560-19568. https://doi.org/10.1609/aaai.v38i17.29928

Du SEM au GEM : redéfinir le métier de référenceur à l'ère des moteurs génératifs

Robert Viseur[1]

1. FWEG, UMONS
Place Warocqué 17, 7000 Mons, Belgique
robert.viseur@umons.ac.be

RÉSUMÉ. À mesure que les moteurs de recherche génératifs transforment la découverte d'information, les stratégies de visibilité héritées du SEO et du SEM deviennent insuffisantes. Cette communication clarifie le périmètre du Generative Engine Optimization et propose le cadre unificateur du Generative Engine Marketing, qui articule optimisation organique, publicité intégrée aux réponses et capacités transactionnelles portées par des agents. Sur la base d'une revue structurée de la littérature, complétée par l'analyse de sources professionnelles, nous construisons une typologie croisant interventions sur le site et hors site, avec des actions réalisées hors moteur ou directement au sein des moteurs. Les résultats mettent en évidence la continuité des leviers SEO, mais aussi des exigences propres au fonctionnement en RAG, telles que la structuration et la justification des contenus. Ils soulignent enfin le rôle décisif de la présence dans des sources tierces privilégiées par chaque agent, ainsi que l'apparition de conventions visant à encadrer le crawling et la réutilisation des contenus par les IA, dont l'efficacité dépend de leur adoption.

ABSTRACT. As generative search engines reshape information discovery, visibility strategies inherited from SEO and SEM are no longer sufficient. This paper clarifies the scope of Generative Engine Optimisation and introduces the unifying framework of Generative Engine Marketing, which connects organic optimisation, advertising embedded in generative answers, and agent-enabled transactional capabilities. Drawing on a structured review of the academic literature, complemented by an analysis of professional sources, we develop a typology that distinguishes on-site and off-site interventions, as well as actions implemented outside the engine or directly within generative engines. The findings highlight continuity with established SEO levers, while also identifying requirements specific to retrieval-augmented generation, notably the need to structure and justify content so that it can be reliably selected and cited. They also stress the decisive role of presence in third-party sources favoured by each agent, alongside the emergence of conventions intended to govern crawling and the re-use of web content by AI systems, whose effectiveness depends on their adoption.

MOTS-CLÉS : SEO, SEM, GEO, LLM.

KEYWORDS: SEO, SEM, GEO, LLM.

1. Problématique

Le secteur de l'e-commerce en Europe connaît une croissance notable depuis de nombreuses années. Sur le plan de la valeur marchande, il dépassait ainsi les 500 milliards de dollars en 2021, après avoir été dopé par les mois de crise de la COVID-19 (Jílková & Králová, 2021). La visibilité, et les revenus, des sites des entreprises dépend en particulier de l'optimisation du référencement (SEO : *Search Engine Optimization*). Celui-ci permet d'améliorer le positionnement des pages du site dans les pages de résultats d'un moteur de recherche. En complément, les entreprises recourent également au SEA (*Search Engine Advertising*), c'est-à-dire l'achat de liens sponsorisés au sein des moteurs de recherche. La combinaison du SEO et du SEA donne le SEM (*Search Engine Marketing*) (Viseur, 2021 ; Dou *et al.*, 2010). Les habitudes de recherche d'information évoluent cependant rapidement depuis la sortie de ChatGPT. Ainsi, près d'un quart des *prompts* dans ChatGPT (21,3%) sont relatifs à une recherche d'information (Chatterji *et al.*, 2025). La recherche de produits commerciaux en est à ses balbutiements (2,1%). Les gestionnaires de sites web doivent dès lors se préparer à l'adaptation de leurs stratégies de référencement au sein des moteurs de recherche génératifs. Une discipline se construit petit à petit, le GEO (*Generative Engine Optimization*), dont les contours restent cependant mal définis (Aggarwal *et al.*, 2024). Notre article explorera dès lors les questions de recherche suivantes. Quelles sont les bonnes pratiques identifiées dans le GEO ? En quoi le GEO se différencie-t-il du SEO ? En quoi consisterait un *Generative Engine Marketing* (GEM) conceptualisé sur le modèle du SEM ?

2. Techniques SEO dans la littérature

Les premiers moteurs de recherche (Yahoo!, Altavista, Lycos), permettant une recherche par mots-clefs, sont apparus au mitant des années quatre-vingt-dix (Amer & Elboghdadly, 2024). Face à l'explosion du nombre de pages, leur importance est allée croissante. La seule pertinence syntaxique n'a rapidement plus suffi à discriminer les pages web dans les résultats, en plus de s'avérer sensible aux manipulations (Amer & Elboghdadly, 2024). Google s'est alors distingué avec son algorithme du Pagerank. Ce dernier permettait d'associer à chaque page un score d'autorité, dépendant de la quantité d'hyperliens (et de leur qualité) renvoyés vers cette page (Amer & Elboghdadly, 2024 ; Cardon, 2013 ; Brin & Page, 1998). Or, les utilisateurs ne regardent en général que la première page de résultats, voire uniquement le Top 3 (Sharma *et al.*, 2019). Le SEO (*Search Engine Optimization*) permet dès lors aux gestionnaires de sites d'optimiser la visibilité dans les résultats organiques des moteurs de recherche (Roumeliotis & Tselikas, 2022), aussi qualifiés de SERP (*Search Engine Results Pages*). Deux types de SEO sont mobilisés : le SEO « *on page* » et le SEO « *off page* » (Shahzad *et al.*, 2020 ; Sharma *et al.*, 2019).

Le SEO « *on page* » porte sur les éléments sous le contrôle du référenceur (Sharma *et al.*, 2019), et en particulier sur l'optimisation du contenu des pages. Le référenceur cherche ainsi à accroître la pertinence syntaxique des contenus. Le

référenceur va dès lors sélectionner des mots-clefs selon un compromis favorable entre une forte popularité (côté consommateurs) et une faible concurrence (côté vendeurs). En cas de forte concurrence, le référenceur peut se rabattre sur des expressions comportant plusieurs mots, généralement moins concurrentielles, suivant ainsi une stratégie de longue traîne (Roumeliotis & Tselikas, 2022 ; Anderson, 2009). Ces mots-clefs vont ensuite être insérés dans le contenu des pages, d'une part, en optimisant leur fréquence (densité), d'autre part, en exploitant la hiérarchie des balises HTML. En particulier, le titre de la page (balise HTML « *<title>* ») va être soigneusement choisi, de même que les titres dans le corps de documents (balises « *<h1>* » à « *<h6>* »). L'originalité des contenus est importante dès lors que les moteurs de recherche tendent à pénaliser les contenus jugés trop similaires (concept de « *duplicate content* »[1]). Le référenceur va aussi renseigner les balises META (« *description* », « *keywords* ») et optimiser leurs URL, notamment par le biais de la réécriture (Roumeliotis & Tselikas, 2022). Une URL statique garantit la stabilité du contenu sous-jacent en plus de permettre l'optimisation du référencement par l'ajout de mots-clefs au sein de l'URL elle-même. Le référencement peut enfin travailler la structure du site de manière à en rendre l'exploration aisée pour les robots d'exploration des moteurs de recherche. Ces précautions sont généralement complétées par des outils permettant de réguler l'activité des robots d'exploration des moteurs de recherche. Il s'agit de la balise META « *robots* » et du protocole d'exclusion des robots (Liu *et al.*, 2025). Le SEO « *off page* » porte sur les éléments hors de contrôle du référenceur (Sharma *et al.*, 2019), et en particulier sur la construction d'un réseau d'hyperliens (« *backlinks* ») pointant vers les pages du site. Le référenceur cherche ainsi à accroître l'autorité de ses contenus aux yeux du moteur de recherche (Cardon, 2013). D'autres techniques viennent compléter cette action : l'encouragement des partages vers les réseaux sociaux et les commentaires au sein des sites ou blogs (Sharma *et al.*, 2019). L'ensemble de ces techniques conventionnelles peut être qualifié de « *White Hat SEO* » (Shahzad *et al.*, 2020), en ce sens qu'elles respectent les recommandations des moteurs de recherche, en particulier Google (Sharma *et al.*, 2019). Des critères secondaires, comme la rapidité de chargement des pages ou la compatibilité avec les terminaux mobiles, se sont par ailleurs rajoutés au fil du temps.

En marge du « *White Hat SEO* » évoluent le « *Black Hat SEO* » (BH) et le « *Grey Hat SEO* » (GH). Les tactiques GH consistent à sur-optimiser les pages, typiquement sur le plan des mots-clefs et des hyperliens, en cohérence avec les critères de pertinence adoptés par les moteurs de recherche (Cardon, 2013). Les tactiques BH recourent quant à elles à des techniques ostensiblement réprouvées par les moteurs de recherche, car considérées comme du *spamindexing* (Sharma *et al.*, 2019 ; Chandra *et al.*, 2015). Premièrement, ces techniques peuvent chercher à jouer sur la pertinence syntaxique. Les mots-clefs peuvent ainsi être multipliés (« *keyword stuffing* ») de manière à en augmenter la densité, soit dans le contenu du document, par exemple par le biais d'un texte invisible, soit dans les URL des pages. Un contenu alternatif optimisé peut également être proposé au robot d'exploration.

[1] Cf. https://developers.google.com/search/blog/2008/09/demystifying-duplicate-content-penalty.

C'est la technique du *cloaking*. Soit le robot d'exploration est détecté (sur base de son nom, c'est-à-dire le *user agent*, ou de son adresse IP fixe, s'ils sont documentés) puis un contenu spécifique lui est envoyé ; soit un contenu optimisé est envoyé par défaut mais un contenu normal est envoyé aux seuls humains par le biais d'une redirection Javascript (que les robots d'exploration ne sont généralement pas capables d'exécuter). Deuxièmement, ces techniques peuvent chercher à jouer sur l'autorité des pages. Un réseau d'hyperliens entrants (*backlinks*) peut être construit à l'aide de pratiques agressives. Ces dernières incluent les fermes de liens (« *link farms* »), permettant d'automatiser les échanges d'hyperliens, et la création de pages satellites (« *doorway pages* »). Dans ce cas, des pages, souvent de faible qualité, sont produites en dehors du site dans le seul but de multiplier les liens entrants. La production de ces pages a été longtemps difficile, et par ailleurs contrée par les dispositifs des moteurs de recherche destinés à lutter contre les contenus dupliqués ou similaires. Cependant, leur production est désormais facilitée par les IA génératives (Viseur, 2023). Au cours de son existence, Google a dès lors été amené à faire évoluer son algorithme, pour améliorer le service rendu aux utilisateurs, mais aussi pour pénaliser les sites recourant à des pratiques abusives pour leur référencement. Peuvent notamment être cités les algorithmes Cassandra en 2003 (lutte contre les fermes de liens) ou les algorithmes Panda, Penguin et Hummingbird (lutte contre les sites de faible qualité) (Patil *et al.*, 2021 ; Chandra *et al.*, 2015).

3. Techniques GEO dans la littérature

Tout comme l'algorithme du Pagerank a été une réponse efficace à la croissance de la taille du Web (Cardon, 2013), les IA génératives proposent une solution à la quantité de résultats renvoyés par les moteurs de recherche et à leur contamination par des contenus de faible qualité exploitant les failles de leurs algorithmes de pertinence (Amer & Elboghdadly, 2024). De plus, les moteurs génératifs introduisent une interaction davantage conversationnelle et contextuelle. En pratique, deux modes d'utilisation doivent cependant être distingués, d'une part, le moteur de réponse, d'autre part, le moteur de recherche génératif. Dans le premier cas, la réponse s'appuie uniquement sur les connaissances embarquées dans le modèle de langage sous-jacent à l'agent conversationnel. Ce modèle est principalement construit à partir de données d'entraînement, issues de la collecte sur le Web (Floridi & Chiriatti, 2020) mais aussi des acquisitions de licences (Stratton, 2025). Dans le second cas, le moteur de recherche génératif collecte des données de contexte avant de synthétiser une réponse avec le modèle de langage sous-jacent. La réponse est dès lors influencée par les données d'entraînement, mais surtout par les données de contexte. Ces dernières dépendent de la transformation du *prompt* en mots-clefs, de la base de données (index) sous-jacente et du modèle de pertinence utilisé (Zhu *et al.*, 2024). Ce mode RAG (*Retrieval-Augmented Generation*) permet d'exploiter la mémoire paramétrique du modèle de langage ainsi que la mémoire non-paramétrique liée à une base de connaissance (Lewis *et al.*, 2020). Le moteur de réponse correspond à ChatGPT sans accès au Web, tel que configuré lors de sa sortie en novembre 2022, tandis que le moteur de recherche génératif correspond plutôt au cas de Perplexity ou de ChatGPT avec accès au Web (p. ex. recherche

simple ou approfondie). Les moteurs de recherche classiques tendent à devenir hybrides comme le montrent l'inclusion de Copilot dans les réponses de Bing ou, dans Google, l'« Aperçu IA » ou le « Mode IA ». Une littérature émergente (par exemple : Aggarwal *et al.*, 2024) explore dès lors l'optimisation du contenu des sites pour les adapter aux moteurs de recherche génératifs. On peut alors parler de GEO (*Generative Engine Optimization*), en complément du SEO (*Search Engine Optimisation*).

Des formats et protocoles spécifiques ont été fabriqués pour contrôler l'action de ces moteurs. Deux familles de techniques peuvent ainsi être distinguées : les techniques de régulation, qui permettent la régulation du comportement des IA sur le site, et les techniques d'optimisation, qui facilitent l'interprétation du contenu par les robots. La première famille comporte deux dispositifs proposés concomitamment à destination des robots éthiques. Le premier dispositif est l'attribut « *noai* » ajouté à la balise HTML META « *robots* ». Cette balise permet ainsi de documenter l'autorisation de réutiliser le contenu par des IA, en plus des options d'indexation (« *(no)index* »), de suivi des hyperliens (« *(no)follow* ») et de mise en cache (« *(no)cache* »). Cette convention est actuellement peu utilisée (Liu *et al.*, 2025). Le second dispositif prévoit l'ajout d'un fichier textuel supplémentaire, nommé « *ai.txt* », permettant de particulariser les droits conférés par type de ressource, en distinguant par exemple le texte, les images et le code source (Li *et al.*, 2025). Ce protocole est actuellement concurrencé par le « *TDM Reservation Protocol* » (TDMRep) proposé par le W3C. La seconde famille distingue les formats spécifiques à la fourniture de contenus structurés. Les fichiers « *llms.txt* » et « *llms-full.txt* »[2] proposent au robot le contenu du site sous forme textuelle, soit un format plus facilement exploitable. Ils offrent ainsi aux agents une information structurée, segmentée, expurgée des balises HTML. La première famille porte donc davantage sur la collecte de données pour l'entraînement, en particulier pour régler la question des conditions d'accès, la seconde, sur la collecte de données de contexte, pour répondre en temps réel à un *prompt*. L'adoption de ces formats spécifiques par les moteurs reste incertaine. Par exemple, Google ne supporte pas l'attribut « *noai* »[3].

En parallèle de ces conventions et formats spécifiques, certains auteurs explorent les techniques d'optimisation des contenus HTML. Aggarwal et ses co-auteurs (2024) mettent en avant l'importance de la structuration des contenus, par exemple sous la forme de questions et de réponses, en plus de la justification des informations comme l'inclusion de citations et de statistiques dans le texte. Chen et ses co-auteurs (2025) confirment l'importance de la structuration des données (p. ex. listes à puce et tableaux comparatifs) mais relèvent par ailleurs l'importance des mentions dans des sources indépendantes (p. ex. presse en ligne et blogs d'experts). Ils insistent cependant sur le fait que les préférences varient d'une IA à une autre, et qu'une stratégie globale pour toutes les IA n'est pas envisageable. De plus, les références utilisées tendent à être localisées, ce qui implique de trouver des relais pour la propagande commerciale dans les différentes langues ciblées. Puerto et ses

[2] Cf. https://github.com/AnswerDotAI/llms-txt.
[3] Cf. https://developers.google.com/search/docs/crawling-indexing/robots-meta-tag.

co-auteurs[4] (2025) insistent pour leur part sur l'importance du SEO. Ce qu'ils soulignent par là, c'est en fait l'influence majeure, sur la réponse générée, du classement des résultats collectés dans les données de contexte.

La littérature a par ailleurs identifié quelques techniques BH : la sur-optimisation des mots-clefs (« *keyword stuffing* »), le *cloaking*, l'injection de *prompt* (« *prompt injection* ») et l'empoisonnement de modèles (« *LLM grooming* »). Comme en SEO BH, le « *keyword stuffing* » consiste à exagérément augmenter la densité en certains mots-clefs. Aggarwal et ses co-auteurs (2024) notent l'inefficacité de cette technique. La collecte de données de contexte est réalisée par des robots, associés à un *user agent* prédéfini, agissant sur des plages d'adresses IP parfois documentées, mais aussi rattachables à des empreintes calculées (« *fingerprinting* »). La détection du robot permet alors l'envoi d'un contenu spécifique, susceptibles d'inclure des instructions trompeuses, par exemple sous la forme d'une injection indirecte de *prompt* (Zychlinski, 2025). Quant au « *LLM grooming* », il consiste à empoisonner délibérément les pages dont le contenu est exploité comme données d'entraînement ou de contexte de manière à orienter les réponses (Alyukov *et al.*, 2025). Cette technique a notamment été discutée dans le contexte des réseaux de désinformation en ligne. Elle présente d'autant plus de chances de fonctionner que le sujet est peu couvert par des sources faisant autorité (« *data voids* »).

4. Analyse des pratiques professionnelles

La littérature professionnelle se développe autour des modalités de référencement au sein des agents conversationnels et des (méta-)moteurs de recherche génératifs. Nous avons constitué un corpus (fourni en Annexe 1) composé de 33 documents, pour un total d'un peu moins de 60.000 mots, ce qui équivaut à 130 pages de texte continu, avec environ 60 % de sources primaires. Ce matériel empirique se compose d'un mélange de documentations techniques, d'annonces officielles et d'analyses (p. ex. presse et experts). Il a fait l'objet d'une analyse axée sur les stratégies, les modèles de revenus, les partenariats, les protocoles et les techniques de référencement. Il nous a permis d'identifier, d'une part, des techniques de référencement organique, hors agent, d'autre part, des techniques de référencement, éventuellement payantes, au sein des agents, sous la forme, soit de liens sponsorisés, soit de capacités transactionnelles.

Les techniques de référencement organiques hors agent cherchent à influencer, d'une part, les données d'entraînement, d'autre part, les données de contexte. Les premières permettent de produire le modèle de langage (LLM) tandis que les secondes alimentent l'IA générative dès lors qu'elle fonctionne en mode RAG (*Retrieval-Augmented Generation*) [EM08]. Trois types d'optimisation peuvent alors être distingués. Le premier type concerne le recours à l'optimisation SEO classique. Une corrélation est ainsi notée entre les résultats issus des moteurs de recherche classiques et ceux issus des moteurs de recherche génératifs. Cependant, il

[4] Ces auteurs ne parlent pas de GEO mais utilisent plutôt l'acronyme C-SEO pour *Conversational Search Engine Optimization*.

ressort que les résultats les mieux classés ne sont pas privilégiés, au contraire des résultats ressortant au-delà de la troisième page de résultats dans les moteurs classiques [EM22]. Cette corrélation peut être attribuée, d'une part, à la plus forte citation des contenus optimisés, d'autre part, à l'utilisation d'API (*Application Programming Interface*) de moteur de recherche par certains agents, à l'image de l'API de Bing utilisée par Perplexity avant que ce dernier ne sorte sa propre Perplexity Search API [EM14]. Cette internalisation est justifiée par des aspects économiques (coûts d'usage des APIs jugés exorbitants) et techniques (p. ex. rapidité de réponses, fraîcheur de l'index et profondeur d'analyse des documents) [EM17]. L'exploitation d'une page nécessite en effet que son contenu puisse être efficacement décomposé (« *chunkable* ») [EM22], ce que l'utilisation des balises HTML (titres « *Hi* », listes, tableaux…) ou de formats de données structurées (schema.org) peut faciliter [EM23]. Le second type concerne l'optimisation dans les données d'entraînement, soit le LLMO (*Large Language Model Optimization*) [EM08]. L'objectif est alors de ressortir parmi les « connaissances » embarquées dans le modèle de langage [EM08]. Le troisième type concerne l'optimisation du référencement parmi les sources utilisées préférentiellement pour alimenter les données de contexte [EM07] [EM08]. Certains distinguent dès lors LLMO (*Large Language Model Optimization*) et GEO (*Generative Engine Optimisation*), les deux pratiques étant chapeautées par l'acronyme GAIO (*Generative Artificial Intelligence Optimisation*) [EM08]. Ces données de contexte sont collectées en vue de préparer la réponse à un *prompt*. Optimiser le référencement dans ces données de contexte suppose d'être cité au sein des références favorisées par chaque moteur [EM07] [EM25]. ChatGPT cite ainsi beaucoup Wikipédia, Reddit, G2 et la presse (p. ex. Forbes, Business Insider et TechRadar) tandis que Perplexity privilégie Reddit puis Youtube ; et Google, YouTube, LinkedIn, Quora et Reddit. Reddit ressort donc comme une plateforme de référence pour la plupart des agents. Cela s'explique notamment par la structuration des données sur ce site et par les offres de licence sur leurs données [EM26]. Reddit participe également à la promotion du protocole Really Simple Licensing (RSL) [EM26]. Ces préférences des moteurs peuvent évoluer au fil du temps [EM07]. Cette technique consistant à influencer les données d'entraînement et de contexte est parfois qualifiée d'ensemencement (« *LLM seeding* ») [EM24].

En matière de liens sponsorisés, Perplexity a dès la fin 2024 testé un dispositif de questions complémentaires sponsorisées (« *sponsored follow-up questions* ») [EM09] monétisées à l'affichage (CPM) [EM12]. Cette expérimentation est conduite avec un nombre limité de partenaires [EM09]. Elle est motivée par l'insuffisance des revenus tirés des abonnements pour atteindre la profitabilité [EM09]. La publicité y complète les offres promotionnelles proposées aux utilisateurs Perplexity Pro [EM12]. Google a également testé la publicité au sein de son Aperçu IA et de son Mode IA [EM11] [EM12], dès lors que le *prompt* présente un angle commercial et en limitant ces tests dans un premier temps aux utilisateurs états-uniens [EM11]. Microsoft fait de même avec Copilot dans Bing [EM10].

En matière de capacités transactionnelles, OpenAI propose Instant Checkout, soit une fonctionnalité d'achats d'articles individuels, actuellement supportée par Etsy et

Shopify [EM01]. Un *prompt* de recherche de produit conduit à « *des résultats organiques et non sponsorisés, classés uniquement par ordre de pertinence* » [EM01]. Si l'article est supporté par Instant Checkout, il s'accompagne d'un bouton « *Buy* ». L'extension des tests hors États-Unis est prévue en 2026. La transaction entre ChatGPT et les commerçants est réglée par le protocole open-source Agentic Commerce Protocol (ACP) [EM02] [EM03]. Ce dernier facilite par ailleurs le paiement du fait de son intégration au service de paiement en ligne Stripe [EM03], qui est co-développeur du protocole ACP [EM01] [EM02]. OpenAI se rémunérerait par une petite commission sur chaque transaction [EM05]. Perplexity propose également un système d'achat au sein de l'agent conversationnel [EM05]. Microsoft propose Copilot Merchant Program [EM05]. Google a également publié son propre protocole open-source baptisé Agent Payments Protocol (AP2) [EM05] [EM20] [EM21]. Celui-ci est une extension des protocoles A2A (*Agent2Agent*) et MCP (*Model Context Protocol*), le premier dédié à l'interopérabilité entre agents et le second à la collecte de données de contexte. Le protocole AP2 permet le paiement depuis un agent, en temps réel ou en différé, avec ou sans humain [EM16]. Il est soutenu par des prestataires de paiements classiques, comme Mastercard et PayPal, ou orientés cryptomonnaies, comme Coinbase et Ethereum Foundation, grâce à l'extension A2A x402 [EM15] [EM18].

Le développement de l'agentique passe aussi par la mise sur le marché de navigateurs intégrant les capacités conversationnelles et agentiques. OpenAI a ainsi publié Atlas [EM29] [EM30] et Perplexity, Comet [EM27] [EM28], tous deux basés sur le navigateur open-source Chromium. Ces navigateurs agentiques combinent les capacités conversationnelles de modèles de langage, leurs capacités de synthèse d'information (liées aux fonctionnalités de recherche simple ou approfondie) et des capacités d'interaction avec des sites web [EM30]. Dans le cas d'Atlas, ces sites web peuvent être rendus pilotables par le biais de conventions comme les balises ARIA (*Accessible Rich Internet Application*) standardisées par le W3C [EM29] ou être assortis d'un connecteur sur mesure compatible MCP [EM31]. Sont notamment ciblées les services de *shopping assistant* permettant l'automatisation progressive des achats, et en particulier la recherche d'offres sur des critères prédéfinis (caractéristiques du produit, délais de livraison, prix…). Le positionnement face aux standards d'agent mobilise aussi les prestataires de paiement comme Stripe ou PayPal [EM04] [EM16].

5. Discussion

Nous pouvons donc distinguer (cf. Tableau 1), parmi les techniques « *off engine* », d'une part, les techniques « *on page* » (c'est-à-dire des régulations et des optimisations réalisées au sein du site à référencer), d'autre part, les techniques « *off page* » (c'est-à-dire les optimisations réalisées en dehors du site). Ces dernières comportent, d'une part, les techniques « *off engine* » (c'est-à-dire des optimisations du référencement, soit dans les données d'entraînement, soit dans les données de contexte), d'autre part, les techniques « *on engine* » (c'est-à-dire des optimisations mises en œuvre au sein des moteurs de recherche génératifs).

Tableau 1. Comparaison SEM et GEM.

	Search Engine Marketing (SEM)	**Generative Engine Marketing (GEM)**
Techniques « *off engine* »	**Search Engine Optimization (SEO) :** Régulation de l'activité : balises META, protocole d'exclusion des robots (REP). Optimisation du référencement : optimisation (HTML) des pages (« *on page* »), optimisation de la structure (maillage) du site (« *on page* ») et stimulation des *backlinks* (« *off page* »). Black Hat SEO : *keyword stuffing* (contenus, URLs), *cloaking*, pages satellites.	**Generative Engine Optimization (GEO) :** Régulation de l'activité : balises META (attribut « noai »), protocole d'exclusion des robots (REP), protocoles dédiés à la configuration des droits (« *ai.txt* », TDMRep, RSL). Optimisation du référencement : optimisation SEO + optimisations spécifiques (GAIO) : (1) ensemencement des sources préférées comme données d'entraînement (LLMO) ou de contexte (GEO) (« *off page* ») et (2) optimisation des pages (GEO) par structuration (balises, schema.org ; QA pros/cons…) et justification (citations, statistiques…) (« *on page* »). Formats dédiés : « llms.txt ». Black Hat GEO : SEO BH (données de contexte alimentées par recherche) ; *cloaking* (données de contexte) ; injection de *prompt* ; empoisonnement des sources préférées (données d'entraînement et de contexte).
Techniques « *on engine* »	**Search Engine Advertising (SEA) :** Liens sponsorisés (CPC, CPM).	**Generative Engine Advertising (GEAd) :** Liens sponsorisés (CPC, CPM). **Generative Engine Agentization (GEAg) :** Protocoles (MCP, A2A, ACP) pour agents (CPA, CPL).

En matière de régulation de l'activité des moteurs de recherche, les moteurs génératifs s'appuient sur les conventions existantes (balise META « *robots* », REP). Les utilisateurs tentent de les étendre (attribut « *noai* ») ou de créer de nouvelles conventions (p. ex. « *ai.txt* », TDMRep et RSL[5]). Dans les deux cas, le succès est conditionné à l'acceptation par les moteurs génératifs. Le précédent créé par l'ACAP (*Automated Access Content Protocol*), censé étendre le REP, illustre la difficulté (Sire, 2015). L'optimisation du référencement peut exploiter les techniques SEO. En effet, les moteurs génératifs s'appuient sur une infrastructure de recherche classique dont les résultats sont ensuite transformés en réponses. Par contre, l'enjeu réside moins dans le positionnement au sein du Top 10 d'une requête que d'un compromis entre un positionnement au sein des premières pages de résultats et d'une structuration appropriée des contenus s'appuyant sur des balises HTML classiques ou des schémas de données structurées. Le référenceur doit ainsi rendre ses pages visibles mais aussi réussir à convaincre le moteur de recherche génératif que le contenu collecté est pertinent au regard du *prompt* qui a initié la ou les recherches. La visibilité d'une marque dans une IA générative (GAIO) peut dépendre, soit de l'optimisation de la présence dans les données d'entraînement du seul modèle de langage (LLMO), soit de l'optimisation de la présence dans les

[5] Un inventaire à jour est proposé sur le site de l'IPTC, cf. https://iptc.org/std/guidelines/data-mining-opt-out/IPTC-Generative-AI-Opt-Out-Best-Practices.pdf.

données de contexte et du respect des critères de sélection des informations utiles lors de la production de la réponse. Dans ce second cas, le contenu HTML doit, d'une part, être structuré (p. ex. balises et attributs sémantiques), d'autre part, inclure des justifications (p. ex. citations et statistiques). De plus, la visibilité dépend de la propagation de la propagande commerciale au sein des sites privilégiés par chaque moteur. Cela permet de distinguer deux stratégies distinctes : une stratégie de référencement de ses propres pages et une stratégie d'ensemencement de pages tierces (ciblant les données d'entraînement et de contexte). Cette stratégie d'ensemencement revient à développer la *Share Of Voice* (SOV) des marques de l'entreprise au sein de ces moteurs. Elle dépend en pratique du domaine et du moteur génératif ciblés. En complément du GEO, vu comme extension du SEO, des formats spécifiques sont proposés par les utilisateurs, comme « *llms.txt* ». Comme pour la régulation, le succès de ces nouvelles conventions dépend de l'acceptation par les moteurs de recherche eux-mêmes et n'est pas garanti. Notre analyse du Top 100 Alexa (dernière version publiée) révèle par ailleurs que 87,4 % de sites proposent un fichier « *robots.txt* », contre 6,3 %, un fichier « *llms.txt* », mais qu'aucun ne supporte « *ai.txt* » ou l'attribut « *noai* ». Wix apparaît par exemple comme un des rares opérateurs de CMS (*Content Management System*) en mode SaaS (*Software as a Service*) à explicitement supporter le protocole « *llms.txt* »[6].

Des techniques de *spam* émergent, parfois inspirées des techniques SEO BH (sur-optimisation des mots-clefs, *cloaking*), parfois spécifiques aux moteurs génératifs (injection de *prompt*, empoisonnement). Relevons que la frontière entre l'ensemencement (« *LLM seeding* ») et l'empoisonnement (« *LLM grooming* ») des sources paraît ténue. Le premier serait plutôt associé aux opérations de relation publique (p. ex. communiqués de presse) et à la production de contenus d'utilisateurs en conformité avec les conditions d'utilisation des plateformes collaboratives tandis que le second remettrait au goût du jour certaines techniques exploitées en SEO BH comme l'utilisation de robots posteurs pour automatiser le relais de la propagande commerciale. Ces techniques sont renforcées par les techniques SEO plus anciennes applicables aux moteurs de recherche classiques, permettant de biaiser la collecte de documents utiles à la production de la réponse.

Parmi les techniques « *on engine* », nous retrouvons les liens sponsorisés (GEAd), classiques dans le domaine de la recherche en ligne puisqu'ils sont à la base du modèle d'affaires de Google. Ce qui est plus nouveau, c'est l'association du moteur génératif avec des protocoles agents (GEAg) tels que MCP, A2A ou ACP, avec une intégration de plus en plus étroite aux navigateurs (p. ex. OpenAI Atlas et Perplexity Comet).

Cette évolution des bonnes pratiques GEO redéfinit les pratiques des référenceurs. Premièrement, les moteurs de recherche génératifs remettent au goût du jour les pratiques « *off page* », liées au relais de la propagande commerciale au travers de sites indépendants. La visibilité des marques, c'est-à-dire leur part de voix (SOV), compte en effet autant (si ce n'est plus) que la visibilité des pages elles-mêmes. Le succès du GEO repose davantage qu'en SEO sur la capacité à diffuser la

[6] Cf. https://support.wix.com/en/article/understanding-your-sites-llmstxt-file.

propagande commerciale au sein des sources privilégiées par ces moteurs. Cela inclut en particulier les plateformes communautaires comme Quora, Reddit et Wikipédia. Deuxièmement, les techniques de référencement « *on page* » évoluent. Ainsi la visibilité des pages sous le contrôle du référenceur nécessite une structuration accrue de l'information, fonction des questionnements des internautes. Cela encourage par exemple la création de nouvelles *landing pages* thématiques. De plus, les stratégies GEO doivent être particularisées par moteur de recherche. L'approche « *one-size-fits-all* », adaptée aux moteurs classiques, ne convient plus pour les moteurs génératifs (Chen *et al.*, 2025). Par ailleurs, les préférences affichées par ces moteurs évoluent, ce qui nécessite un travail d'ajustement en continu. Compte tenu des parts de marché actuelles, surtout significatives pour ChatGPT[7], il est cependant envisageable de ne cibler que ChatGPT.

Depuis plus d'une décennie, le Web s'éloigne d'une conception ouverte où les services seraient consommés depuis le navigateur, ce qui se traduit par le développement des *apps* et des jardins clos (Paterson, 2012). Cette tendance s'est notamment concrétisée par des plateformes comme Facebook et Amazon. Ces plateformes se caractérisent par deux processus concomitants, permis par leur caractère programmable (disponibilité d'API) : la décentralisation des fonctionnalités et la recentralisation des données. Les moteurs de recherche génératifs ont besoin de données structurées pour produire des réponses pertinentes (Chen *et al.*, 2025). Au-delà des techniques d'optimisation « *on page* », les protocoles d'agent comme MCP viennent combler ce besoin en données structurées et contribuent à la programmabilité des services, en particulier dans le contexte des activités d'ecommerce (p. ex. comparaison des offres). Au-delà de la seule recherche d'information, en transformant le navigateur en assistant IA (p. ex. Atlas et Comet), les producteurs d'IA génératives se positionnent comme de nouvelles plateformes. Elles offrent ainsi un point d'accès unique aux services en ligne. Cette évolution impacte le métier des référenceurs mais perturbe surtout sensiblement les rapports de force établis depuis plusieurs années.

Premièrement, l'autonomie des consommateurs est affectée. En effet, l'agent joue un rôle de prescripteur en recherchant des solutions disponibles. Il se pose comme un assistant à la prise de décision, par exemple en évaluant les offres puis en présentant un classement (Chen *et al.*, 2025) voire en automatisant les achats (*shopping assistant*). D'une part, ce mode d'interaction introduit de nouvelles vulnérabilités. De nouveaux mécanismes de lutte contre la fraude [EM04] sont nécessaires puisque l'acte d'achat repose substantiellement sur les actions de l'IA. Une simulation réalisée par Microsoft dans son environnement Magentic Marketplace montre ainsi le bon fonctionnement des grands modèles de langage dans un cadre idéal mais de faiblesses importantes face à des tentatives de manipulation incluant la saturation en information des agents (Bansal *et al.*, 2025). D'autre part, la conversation devient l'interface, soit une tendance inaugurée par les assistants vocaux tels que Siri ou Amazon Alexa (Amer & Elboghdadly, 2024). Le contrôle laissé à l'utilisateur est dès lors réduit (Vachaudez & Koubi, 2023), ce qui s'explique par exemple par le fait que l'agent réduit le spectre des alternatives

[7] Cf. https://gs.statcounter.com/ai-chatbot-market-share.

présentées (Epstein *et al.*, 2022) ou que l'initiative est davantage reportée sur le système (Bérubé *et al.*, 2024). La capacité à influencer l'agent devient donc critique pour le vendeur, et le référenceur en charge de la visibilité des produits.

Deuxièmement, dans le cadre d'activités de commerce en ligne, les agents conversationnels viennent contester le rôle de *gatekeeper* joué traditionnellement par les plateformes comme Amazon [EM05]. Rappelons que, lors d'une recherche de produit, 55 % des internautes utiliseraient Amazon contre 28 % Google (Galloway, 2018). Cette position permet par exemple à Amazon la mise en œuvre de badges algorithmiques (p. ex. « *Amazon's Choice* ») ainsi que l'auto-préférence (*self-preferencing*), une pratique contre laquelle le DMA (*Digital Markets Act*) européen entend d'ailleurs lutter. Elle lui offre aussi un revenu complémentaire lié à la commercialisation de liens sponsorisés [EM33]. Les moteurs de recherche génératifs pourraient réduire rapidement le pouvoir de prescription des plateformes existantes (p. ex. Amazon Marketplace) en facilitant la comparaison automatisée des offres. C'est d'ailleurs une des fonctionnalités proposées par Perplexity et OpenAI au travers, d'une part, de leurs agents, d'autre part, de leurs navigateurs agentiques. Si dans un premier temps cette dynamique semble favorable au consommateur, à moyen ou long terme, les éditeurs de ces systèmes pourraient eux-mêmes user de ce pouvoir prescriptif pour privilégier les offres qui leur sont le plus profitables.

Cette menace sur les positions établies devrait amener les plateformes ecommerce dominantes à réagir. Premièrement, elles devraient elles-mêmes monter en compétences sur les techniques permettant d'influencer ces agents, au travers des pages HTML ou des protocoles mis à disposition. D'ailleurs, les entreprises technologiques luttent actuellement pour l'adoption de leurs technologies en recourant à des stratégies de diffusion de standards sous licence libre [EM05], une pratique courante dans le secteur informatique (Adatto, 2013). Deuxièmement, elles devraient s'opposer aux nouveaux entrants. La plainte d'Amazon contre Perplexity constitue une première réponse, judiciaire, pour freiner le développement des *shopping assistants*. Une seconde réponse est plus technique. Ainsi, ni Amazon ni eBay n'adhèrent à ACP ou A2P. Soucieuses de défendre leur rôle de *gatekeeper*, les deux plateformes privilégient leurs propres API et leurs outils génératifs internes. Amazon, par exemple, propose l'assistant Rufus, entraîné sur des données internes (catalogues, commentaires…) [EM20] (Chilimbi, 2024), tandis qu'eBay a investi, quoique plus timidement, dans sa propre famille de LLM baptisée LiLiuM (Herold *et al.*, 2024). Cela signifie, pour le référenceur, la nécessité de particulariser la stratégie de référencement à des plateformes distinctes, incluant les moteurs de recherche classiques, les plateformes ecommerces (jouant un rôle de moteur de recherche de produits) et les agents conversationnels, qu'ils soient utilisés comme simple moteur de recherche génératif ou comme *shopping assistant* autonome.

5.1. Limitations et perspectives

Notre recherche propose un premier effort de synthèse en matière d'optimisation du référencement dans les moteurs de recherche génératifs. Elle repose sur un

mélange de sources académiques et de matériel professionnel ou journalistique permettant d'explorer les pratiques actuellement émergentes et d'en comprendre la rationalité. Cependant, elle ne propose pas de métriques (p. ex. nombre d'impressions du site, taux d'inclusion du site, position du site dans le résumé, part de voix du site et taux de couverture du site). De telles métriques permettraient notamment de justifier et hiérarchiser les techniques proposées au sein du tableau récapitulatif (cf. Tableau 1).

Les moteurs de recherche génératifs précèdent la génération de la réponse d'une phase de recherche de pages, soit à l'aide d'une API de moteur de recherche classique (p. ex. Bing Search API), soit à l'aide d'un index propre (p. ex. Perplexity Search API). Cela signifie que ressortir dans la synthèse produite par un moteur de recherche génératif nécessite d'être préalablement *shortlisté* à l'issue de la recherche de pages pertinentes. Quelles sont les techniques SEO (WH, GH et BH) permettant d'influencer les résultats de ces nouvelles plateformes d'indexation associées aux moteurs génératifs (p. ex. Perplexity) ? Présentent-elles des spécificités comparativement aux moteurs de recherche traditionnels (Google, Bing) ? Par exemple, la décomposabilité du contenu d'une page pourrait être retenue comme un signal de qualité pris en compte favorablement par le modèle de pertinence.

Les efforts en matière de Web sémantique ont conduit à la création de standards, permettant l'annotation d'un document HTML sur base d'un vocabulaire prédéterminé (Meusel *et al.*, 2014). Citons par exemple les microformats, le RDFa et les microdata. Chen et ses co-auteurs (2025) évoquent ainsi l'usage des schémas de données structurées couverts par la communauté Schema.org (RDFa, microdata, JSON-LD) pour répondre aux besoins des IA en données structurées de contexte. Son efficacité sur le plan du GEO n'est cependant pas évaluée par les auteurs, alors que ces formats sont exploitables en SEO, car pris en compte par les moteurs de recherche[8]. Or, Schema.org fournit notamment les types « *Product* » et « *Offer* », ce qui permet de décrire un produit, son prix, sa devise, sa disponibilité, son état ou encore ses identifiants (p. ex. GTIN, ASIN et SKU). Comparée aux protocoles dédiés aux agents ecommerce, cette approche permet de privilégier des outils associés aux standards du Web ouvert. Leur utilisation dans une boutique en ligne est-elle corrélée avec une bonne visibilité au sein d'un moteur de recherche génératif ?

6. Conclusion

Cette recherche met en évidence un basculement dans les pratiques de référencement à l'ère des moteurs de recherche génératifs. Alors que le SEO (*Search Engine Optimization*) visait à optimiser la visibilité des pages dans des systèmes fondés sur des index lexicaux et des modèles de pertinence statiques, le GEO (*Generative Engine Optimization*) cherche désormais à influencer la manière dont

[8] Voir par exemple Google : https://developers.google.com/search/docs/appearance/structured-data/intro-structured-data et https://developers.google.com/search/docs/appearance/structured-data/merchant-listing.

les agents conversationnels, fondés sur des LLM, collectent, sélectionnent et synthétisent l'information, voire agissent sur base des consignes fournies par l'utilisateur. Cette recherche nous a tout d'abord permis d'identifier un ensemble de bonnes pratiques, en matière d'optimisation, soit « *on page* », soit « *off page* », et, dans ce second cas, « *on engine* » ou « *off engine* ». Elle met ensuite en évidence la remise au goût du jour des pratiques de diffusion de la propagande commerciale de manière à augmenter la part de voix (*share of voice*) au sein des sources privilégiées par les moteurs de recherche pour l'acquisition de données d'entraînement ou de contexte. Elle nous a enfin permis de définir les caractéristiques du GEM en distinguant les techniques liées à l'optimisation des contenus (GEO), aux liens sponsorisés (GEAd) et aux capacités transactionnelles (GEAg). Cette dynamique affecte, premièrement, les pratiques des référenceurs confrontés à de nouveaux outils aux comportements hétérogènes, deuxièmement, l'autonomie des utilisateurs confrontés à des interfaces conversationnelles fonctionnant en boîte noire, troisièmement, la position dominante des plateformes.

7. Références

Adatto, L. (2013). Standards ouverts et implémentations FLOSS: vers un nouveau modèle synergique de standardisation promu par l'industrie du logiciel. Terminal. Technologie de l'information, culture & société, (113-114), 137-172. https://doi.org/10.4000/terminal.293.

Aggarwal, P., Murahari, V., Rajpurohit, T., Kalyan, A., Narasimhan, K., & Deshpande, A. (2024). GEO: Generative engine optimization. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 5-16). https://doi.org/10.1145/3637528.3671900.

Alyukov, M., Makhortykh, M., Voronovici, A., & Sydorova, M. (2025). LLMs grooming or data voids? LLM-powered chatbot references to Kremlin disinformation reflect information gaps, not manipulation. Harvard Kennedy School Misinformation Review. https://doi.org/10.7910/DVN/LHGU1O.

Amer, E., & Elboghdadly, T. (2024). The End of the Search Engine Era and the Rise of Generative AI: A Paradigm Shift in Information Retrieval. In 2024 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC) (pp. 374-379). IEEE. https://doi.org/10.1109/MIUCC62295.2024.10783559.

Anderson, C. (2009). La longue traîne. Comment Internet a bouleversé les lois du commerce. Pearson. ISBN : 978-2744063855.

Bansal, G., Hua, W., Huang, Z., Fourney, A., Swearngin, A., Epperson, W., ... & Amershi, S. (2025). Magentic Marketplace: An Open-Source Environment for Studying Agentic Markets. arXiv preprint arXiv:2510.25779. https://doi.org/10.48550/arXiv.2510.25779.

Bérubé, C., Nißen, M., Vinay, R., Geiger, A., Budig, T., Bhandari, A., ... & Kocaballi, A. B. (2024). Proactive behavior in voice assistants: A systematic review and conceptual model. Computers in Human Behavior Reports, 14, 100411. https://doi.org/10.1016/j.chbr.2024.100411.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. Computer networks and ISDN systems, 30(1-7), 107-117. https://doi.org/10.1016/S0169-7552(98)00110-X.

Cardon, D. (2013). Dans l'esprit du PageRank: Une enquête sur l'algorithme de Google. Réseaux, (1), 63-95. https://doi.org/10.3917/res.177.0063.

Chandra, A., Suaib, M., & Beg, R. (2015). Google Search Algorithm updates against web spam. Department of Computer Science & Engineering, Integral University, 3(1), 1-10.

Chatterji, A., Cunningham, T., Deming, D. J., Hitzig, Z., Ong, C., Shan, C. Y., & Wadman, K. (2025). How people use ChatGPT (No. w34255). National Bureau of Economic Research. https://doi.org/10.3386/w34255.

Chen, M., Wang, X., Chen, K., & Koudas, N. (2025). Generative Engine Optimization: How to Dominate AI Search. arXiv preprint arXiv:2509.08919. https://doi.org/10.48550/arXiv.2509.08919.

Chilimbi, T. (2024). How We Built Rufus, Amazon's AI-Powered Shopping Assistant. IEEE Spectrum. https://spectrum.ieee.org/amazon-rufus.

Dou, W., Lim, K. H., Su, C., Zhou, N., & Cui, N. (2010). Brand positioning strategy using search engine marketing. MIS quarterly, 261-279. https://doi.org/10.2307/20721427.

Epstein, R., Lee, V., Mohr, R., & Zankich, V. R. (2022). The Answer Bot Effect (ABE): A powerful new form of influence made possible by intelligent personal assistants and search engines. PloS one, 17(6), e0268081. https://doi.org/10.1371/journal.pone.0268081.

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. Minds and machines, 30(4), 681-694. https://doi.org/10.1007/s11023-020-09548-1.

Galloway, S. (2018). The four - Le règne des quatre : la face cachée d'Amazon, Apple, Facebook et Google. Quanto. ISBN : 978-2889152469.

Herold, C., Kozielski, M., Ekimov, L., Petrushkov, P., Vandenbussche, P. Y., & Khadivi, S. (2024). LiLiuM: eBay's Large Language Models for e-commerce. arXiv preprint arXiv:2406.12023. https://doi.org/10.48550/arXiv.2406.12023.

Jílková, P., & Králová, P. (2021). Digital consumer behaviour and ecommerce trends during the COVID-19 crisis. International Advances in Economic Research, 27(1), 83-85. https://doi.org/10.1007/s11294-021-09817-4.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems, 33, 9459-9474. https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

Li, Y., Song, W., Zhu, B., Gong, D., Liu, Y., Deng, G., ... & Xue, J. (2025). ai. txt: A Domain-Specific Language for Guiding AI Interactions with the Internet. arXiv preprint. https://doi.org/10.48550/arXiv.2505.07834.

Liu, E., Luo, E., Shan, S., Voelker, G. M., Zhao, B. Y., & Savage, S. (2025). Somesite I Used To Crawl: Awareness, Agency and Efficacy in Protecting Content Creators From AI Crawlers. IMC 2025, Madison (USA). https://doi.org/10.48550/arXiv.2411.15091.

Meusel, R., Petrovski, P., Bizer, C. (2014). The WebDataCommons Microdata, RDFa and Microformat Dataset Series. In: Mika, P., et al. The Semantic Web – ISWC 2014. ISWC 2014. Lecture Notes in Computer Science, vol 8796. Springer, Cham. https://doi.org/10.1007/978-3-319-11964-9_18.

Paterson, N. (2012). Walled gardens: the new shape of the public Internet. In Proceedings of the 2012 iConference (pp. 97-104). https://doi.org/10.1145/2132176.2132189.

Patil, A., Pamnani, J., & Pawade, D. (2021). Comparative Study Of Google Search Engine Optimization Algorithms: Panda, Penguin and Hummingbird. In 2021 6th International Conference for Convergence in Technology (I2CT) (pp. 1-5). IEEE. https://doi.org/10.1109/I2CT51068.2021.9418074.

Puerto, H., Gubri, M., Green, T., Oh, S. J., & Yun, S. (2025). C-SEO Bench: Does Conversational SEO Work?. arXiv preprint arXiv:2506.11097. https://doi.org/10.48550/arXiv.2506.11097.

Roumeliotis, K. I., & Tselikas, N. D. (2022). An effective SEO techniques and technologies guide-map. Journal of web engineering, 21(5), 1603-1649. https://doi.org/10.13052/jwe1540-9589.21510.

Shahzad, A., Jacob, D. W., Nawi, N. M., Mahdin, H., & Saputri, M. E. (2020). The new trend for search engine optimization, tools and techniques. Indonesian Journal of Electrical Engineering and Computer Science, 18(3), 1568-1583. http://doi.org/10.11591/ijeecs.v18.i3.pp1568-1583.

Sharma, D., Shukla, R., Giri, A. K., & Kumar, S. (2019, January). A brief review on search engine optimization. In 2019 9th international conference on cloud computing, data science & engineering (confluence) (pp. 687-692). IEEE. https://doi.org/10.1109/CONFLUENCE.2019.8776976.

Sire, G. (2015). Inclusion exclue: le code est un contrat léonin. Réseaux, 189(1), 187-214. https://doi.org/10.3917/res.189.0187.

Stratton, M. (2025). Market-Based Licensing for Publishers' Works Is Feasible. Big Tech Agrees. Colum. JL & Arts, 48, 434. https://doi.org/10.52214/jla.v48i4.13925.

Vachaudez, A., & Koubi, C. (2023). OK Google:"Why do users of voice assistants maintain their use of technology over time?". Décisions Marketing, 112(4), 177-197. https://doi.org/10.3917/dm.112.0027.

Viseur, R. (2023). Éthique du dropshipping SEO à l'ère des IA génératives. Management & Datascience, 7(4). https://doi.org/10.36863/mds.a.25311.

Viseur, R. (2021). Du tracking, des contre-mesures et de leur efficacité dans la publicité ciblée. Revue ouverte d'ingénierie des systèmes d'information, 2(1). https://doi.org/10.21494/ISTE.OP.2021.0603.

Zhu, Y., Yuan, H., Wang, S., Liu, J., Liu, W., Deng, C., Chen, H., Liu, Z., Dou, Z. & Wen, J.-R. (2024). Large Language Models for Information Retrieval: A Survey, arXiv. https://doi.org/10.48550/arXiv.2308.07107.

Zychlinski, S. (2025). A Whole New World: Creating a Parallel-Poisoned Web Only AI-Agents Can See. arXiv preprint arXiv:2509.00124. https://doi.org/10.48550/arXiv.2509.00124.

8. Annexe 1 : matériel empirique

ID	URL	Mots-clefs
EM01	https://openai.com/index/buy-it-in-chatgpt/	OpenAI, ChatGPT, ACP
EM02	https://github.com/agentic-commerce-protocol/agentic-commerce-protocol	ACP
EM03	https://www.agenticcommerce.dev/	ACP
EM04	https://stripe.com/fr-be/guides/agentic-commerce	commerce agentique, Stripe, MCP, Perplexity
EM05	https://techcrunch.com/2025/09/29/openai-takes-on-google-amazon-with-new-agentic-shopping-system/	OpenAI, ACP, Perplexity, Microsoft, Copilot Merchant Program, Google, Agent Payments Protocol (AP2)
EM06	https://www.journaldunet.com/retail/1544935-agentic-commerce-quand-l-ia-transforme-la-conversation-en-tunnel-d-achat/	ACP, GEO
EM07	https://www.azoma.ai/insights/wikipedia-chatgpt-citations-and-traffic-growth	ChatGPT, Perplexity, Google
EM08	https://otterly.ai/research/OtterlyAI_Generative_Engine_Optimization_Guide.pdf	GEO, LLMO
EM09	https://searchengineland.com/perplexity-begins-testing-ads-448277	Perplexity, liens sponsorisés
EM10	https://ppcnewsfeed.com/ppc-news/2025-02/ads-in-copilot-results-in-bing/	Microsoft advertising
EM11	https://www.theverge.com/2024/10/3/24260637/googles-ai-overview-ads-launch	Google, liens sponsorisés
EM12	https://www.businessinsider.com/inside-perplexity-ai-advertising-pitch-sponsored-questions-perks-2025-6	Perplexity, liens sponsorisés
EM13	https://chatgpt.com/fr-FR/merchants/	OpenAI, ACP
EM14	https://www.perplexity.ai/fr/hub/blog/introducing-the-perplexity-search-api	Perplexity, moteur de recherche, API
EM15	https://cloud.google.com/blog/products/ai-machine-learning/announcing-agents-to-payments-ap2-protocol	Google, AP2
EM16	https://developer.paypal.com/community/blog/PayPal-Agent-Payments-Protocol/	PayPal, Google AP2
EM17	https://www.perplexity.ai/api-platform/resources/architecting-and-evaluating-an-ai-first-search-api	Perplexity, recherche, API
EM18	https://a2aprotocol.ai/ap2-protocol	A2A, A2P
EM19	https://www.amazon.com/Rufus/	Amazon, Rufus
EM20	https://ap2-protocol.org/	A2P
EM21	https://github.com/google-agentic-commerce/AP2	A2P
EM22	https://www.semrush.com/blog/ai-search-seo-traffic-study/	GEO
EM23	Charlot, C. (2025). Le marketing digital à l'ère de l'IA. 7 règles d'or pour préparer sa marque au GEO. Trends Tendance, 11 septembre 2025.	GEO
EM24	https://backlinko.com/llm-seeding	LLM seeding
EM25	https://www.tryprofound.com/blog/ai-platform-citation-patterns	GEO
EM26	https://www.cjr.org/analysis/reddit-winning-ai-licensing-deals-openai-google-gemini-answers-rsl.php	Reddit, licences
EM27	https://www.perplexity.ai/fr/hub/blog/introducing-comet	Perplexity, Comet, agent, navigateur
EM28	https://www.perplexity.ai/fr/hub/blog/shopping-that-puts-you-first	Perplexity, shopping assistant
EM29	https://openai.com/fr-FR/index/introducing-chatgpt-atlas/	OpenAI, Atlas, agent, navigateur
EM30	https://openai.com/fr-FR/index/introducing-chatgpt-agent/	OpenAI, agent
EM31	https://help.openai.com/en/articles/11487775-connectors-in-chatgpt	OpenAI, agent, connecteurs, MCP
EM32	https://www.theguardian.com/technology/2025/nov/18/amazon-vs-perplexity-the-ai-agent-war-has-arrived	Perplexity, Amazon, agent war
EM33	https://www.perplexity.ai/fr/hub/blog/bullying-is-not-innovation	Perplexity, Amazon, plainte

Détection de références bibliographiques hallucinées ou rétractées : bibCheck

Léo Gaillard[1], Victoria Meneghel[1], Pascal Cuxac[1], Guillaume Cabanac[2, 3]

1. *Institut de l'information scientifique et technique (Inist-CNRS, UAR76), 2 rue Jean Zay, F-54500 Vandoeuvre-lès-Nancy prenom.nom@inist.fr*

2. *Institut de Recherche en Informatique de Toulouse (IRIT, UMR 5505), Université de Toulouse, 118 route de Narbonne, F-31062 Toulouse guillaume.cabanac@univ-tlse3.fr*

3. *Institut Universitaire de France (IUF), 1 rue Descartes, F-75005 Paris*

RESUME. *Les articles scientifiques frauduleux ou générés par IA sont de plus en plus nombreux, s'accompagnant de références erronées ou rétractées. Nous présentons l'algorithme bibCheck pour détecter les références problématiques dans les bibliographies. Il automatise la vérification de l'existence des DOI et l'analyse des éléments clés d'une référence bibliographique pour identifier celles qui sont hallucinées, introuvables ou rétractées (grâce au Problematic Paper Screener). Pour vérifier l'existence d'une référence dans Crossref ou DataCite, une similarité partielle floue est calculée entre les différents champs bibliographiques. Les résultats atteignent une macro F-mesure de 0,89. L'algorithme mis en production par l'Inist-CNRS est simple à utiliser, permettant de limiter la propagation de la « mauvaise science ».*

ABSTRACT. *AI-generated articles are booming, and they come with erroneous or retracted references. This article introduces the bibCheck algorithm to detect problematic bibliographic entries. It combines DOI verification via Crossref or DataCite, and metadata analysis to identify references that are hallucinated, nonexistent, or retracted (using the Problematic Paper Screener). References are checked against Crossref and DataCite by computing a partial fuzzy similarity between various bibliographic fields. The results yield a macro F-score of 0.89. The algorithm in production at Inist-CNRS is easy to use and helps to reduce the spread of 'bad science.'*

MOTS-CLÉS : *référence bibliographique, détection d'anomalie, DOI, LLM, hallucination*

KEYWORDS: *bibliographic reference, anomaly detection, DOI, LLM, hallucination*

1. Introduction

Avec l'évolution des méthodes de génération de texte, notamment depuis l'introduction des grands modèles de langues (LLM), les communautés scientifiques sont confrontées à un nombre croissant d'articles générés par l'IA. Le nombre d'articles scientifiques partiellement générés par l'IA varie selon le domaine d'étude, mais pourrait atteindre 22 % des articles sur les *computational sciences* entre 2020 et 2024 (Liang *et al.*, 2025).

L'évolution des méthodes génératives pose un autre problème : celui des fausses références, générées de façon probabiliste par les LLM, appelées « références hallucinées » (Cossio, 2025; Naddaf et Quill, 2026). Une hallucination y est décrite comme une génération de contenu plausible mais factuellement incorrect, incohérent ou entièrement fabriqué. L'ampleur de ce phénomène a conduit certaines maisons d'édition à mettre en place des procédures pour contrôler la validité des bibliographies des manuscrits soumis. Par exemple, Springer Nature développe une IA pour vérifier l'intégrité des bibliographies[1] mais des lecteurs ont détecté des références hallucinées, y compris dans un ouvrage sur l'éthique de l'IA générative : un cas ironique commenté par le *Times* qui mentionne bibCheck[2] en décembre 2025.

La présence de références hallucinées dans une bibliographie peut entraîner la rétractation d'un article car elle traduit un défaut de soin, voire de probité lorsqu'elle révèle une méconduite de la part des auteurs. Sur ce fondement, une des premières rétractations date d'avril 2024 : la revue *PLOS One* avait publié un article dont 18 références sur 76 n'existent pas[3]. D'autres motifs peuvent conduire à la rétractation : données erronées ou falsifiées, faute involontaire ou encore présence de références bibliographiques rétractées (cas des rétractations en cascade). La rétractation peut être à l'initiative de l'auteur ou des éditeurs scientifiques ; elle est opérée par la maison d'édition qui a publié l'article.

Plusieurs ressources s'efforcent de recenser et de signaler les publications rétractées pour éviter que de futurs articles ne s'appuient sur ces références reconnues comme étant non fiables. Le Problematic Paper Screener (PPS, Cabanac *et al.*, 2022; Cabanac, 2024a,b) est une contribution de la recherche publique française visant à collecter, entres autres, ces articles rétractés. Le PPS opère la synthèse[4] des sources principales en la matière : la Retraction Watch Database accessible via Crossref (Lammey et Oransky, 2024), Crossmark et Pubmed. Globalement, le nombre d'articles rétractés croît et a atteint un pic en 2023 à près de 20 000 articles rétractés. La proportion a presque doublé entre 2000 et 2023, passant de 1 ‰ à 2 ‰ (Van Noorden, 2023). Les rétractations croissent davantage dans certaines régions du globe (Cabanac *et al.*, 2023) et pour certains organismes de recherche (Van Noorden, 2025). Un des facteurs explicatifs a

1. Communiqué de presse du 7 avril 2025 : https://group.springernature.com/gp/group/media/press-releases/new-research-integrity-ai-tool/27769148
2. https://pubpeer.com/publications/FDF817E4E2B83B0D1E06DBBDA01844
3. https://pubpeer.com/publications/7F01171F32B0421DF37EE8D1B49C04#5
4. https://www.irit.fr/~Guillaume.Cabanac/problematic-paper-screener/annulled

trait au recours à des *paper mills* qui vendent à des auteurs peu scrupuleux des articles fabriqués à façon (Else et Van Noorden, 2021). Ainsi, identifier les références rétractées dans les manuscrits soumis et dans des publications à ré-évaluer est de plus en plus nécessaire.

C'est dans ce contexte qu'a émergé le besoin d'un algorithme de vérification de références bibliographiques. D'une part, pour éviter la citation d'articles rétractés. D'autre part, pour repérer les fausses références et a fortiori les articles entièrement générés par IA. En effet, un tel article comporte des références générées par l'IA, comme discuté par de nombreux auteurs (Alkaissi et McFarlane, 2023; Gravel *et al.*, 2023; Eiko, 2023) et étudié par Chelli *et al.* (2024). L'augmentation des citations générées par les LLM sur la plateforme de preprint arXiv a également été documentée (Tramèr, 2025). Ce phénomène ne se limite pas à cette archive ouverte : des « *halluCitations* » ont également été observées dans les actes de conférences internationales, telles que ceux édités par l'*Association for Computational Linguistics (ACL)* (Sakai *et al.*, 2026).

Pour apporter une réponse à ces problèmes, nous avons conçu l'algorithme bibCheck qui détecte les références hallucinées ainsi que les références rétractées. Il prend en entrée une référence bibliographique et éprouve sa validité. Le traitement consiste à vérifier sa présence auprès des agences d'enregistrement de DOI (*Digital Object Identifier*) Crossref et DataCite[5]. Il s'agit des deux principales sources de métadonnées indexant la majorité des articles scientifiques avec DOI. La comparaison opérée détermine si la référence existe, est introuvable (potentiellement hallucinée) ou contient des erreurs. Les références validées car existantes sont enfin confrontées aux DOI identifiés comme rétractés par le PPS.

La figure 1 illustre une même référence bibliographique rédigée conformément à deux normes bibliographiques (APA et Vancouver). Elle montre le découpage de la référence en cinq composantes : les auteurs ou autrices, le titre, la date, la source et le DOI. Ces champs sont exploités par l'algorithme bibCheck.

L'algorithme bibCheck est utilisable via une API ou via le site TDM Factory qui sert d'interface à plusieurs services web (Gaillard *et al.*, 2026). Le déploiement technique de bibCheck est détaillé dans la section 4.

2. État de l'art : détection d'articles erronés voire frauduleux

Avant l'émergence des IA génératives, des algorithmes produisaient déjà des contenus textuels intégrés dans des publications scientifiques bien que dénués de sens. C'est le cas de SCIgen ou encore Mathgen (Ball, 2005; Cabanac et Labbé, 2021). Pour détecter les textes intégralement générés, des méthodes probabilistes ont été mises au point (Labbé *et al.*, 2016). Elles ne sont pas applicables aux nouvelles formes de fraude (notamment celles utilisant des LLM) qui ont nécessité des approches différentes. Par exemple, Cabanac *et al.* (2021a) ont découvert des plagiats par copier–paraphraser–

5. https://www.doi.org/the-community/existing-registration-agencies

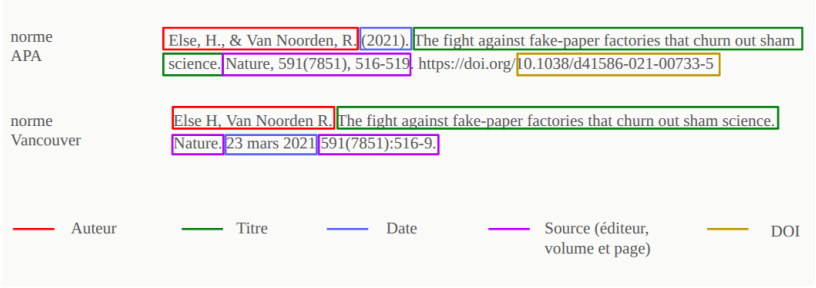


FIGURE 1 : Une référence bibliographique rédigée conformément à la norme APA et à la norme Vancouver (cf. le service DOI citation formatter). Les rectangles colorés délimitent les champs repérés et utilisés par l'algorithme bibCheck.

coller produisant des *tortured phrases*, expressions produites par remplacement de mots par des synonymes. Clausse *et al.* (2024) ont appliqué la même logique aux abréviations torturées lorsque les initiales ne correspondent pas. Par exemple, il est anormal de trouver *"Back spread (BP)"* au lieu de *"Back propagation (BP)"* dans un article et cet indice suggère une méconduite (O'Grady, 2024).

Dans certains cas, les données ou les résultats sont falsifiés. Labbé *et al.* (2020) ont fouillé des textes des publications biomédicales pour identifier des séquences de nucléotides erronées. Les citations constituent un autre type de texte difficile à générer parfaitement par l'IA : en effet, citer correctement et précisément un article et l'associer à la bonne idée nécessite une analyse fine de ce dernier. C'est pourquoi des travaux visent à détecter les références « mal citées » (Liu *et al.*, 2024). Tant pour les références fausses que pour les références mal citées, les IA grand public sont encore peu performantes dans la génération de références authentiques. Contrairement aux méthodes fondées sur la détection de phrases générées par l'IA, les modèles ne semblent pas s'améliorer de manière significative sur les benchmarks à l'heure actuelle, comme mentionné dans (Chelli *et al.*, 2024).

Comme évoqué précédemment, le PPS s'appuie sur plusieurs sources (dont la base de données de Retraction Watch). L'algorithme bibCheck utilise le PPS, avec un double objectif. Le principal consiste à signaler les publications qui ont déjà été rétractées afin d'empêcher la propagation de mauvaises pratiques scientifiques (Cabanac, 2024a). Malgré leur retrait, ces articles rétractés continuent d'être cités et peuvent « contaminer » d'autres articles qui les citent. Le deuxième objectif consiste à identifier les références qui ne sont pas enregistrées par les maisons d'édition dans Crossref ou DataCite. C'est notamment le cas des références générées par IA, qui fabrique par assemblage probabiliste les champs bibliographiques.

Depuis la mise en production[6] de bibCheck en décembre 2024, des approches similaires ont vu le jour. Guenci *et al.* (2025) proposent un processus pour vérifier une référence bibliographique. Les critères de validations sont plus complexes que dans bibCheck, mais ni les références hallucinées ni les rétractées ne sont détectées. Il s'agit plutôt de vérifier la complétude des champs d'une référence. Le logiciel CERCA[7] publié en janvier 2026 est plus interactif. Comme bibCheck, CERCA traite un PDF et structure automatiquement les références avant de les vérifier. Ce logiciel développé en Java permet de corriger les références mal extraites et de les vérifier de façon interactive. C'est un avantage dans le cas de références incorrectement extraites : CERCA permet leur modification avant de relancer les vérifications de manière unitaire. Cette permissivité est utile pour des utilisateurs réguliers qui ont un besoin fréquent d'un tel outil. A contrario, bibCheck est moins interactif mais a l'avantage d'être utilisable sans installation, directement depuis le web. Un autre algorithme permettant de détecter les fausses références est mentionné dans le blog d'Aksenfeld (2025), mais aucun article scientifique associé n'a été trouvé. Le code de cet outil est accessible en ligne[8]. Enfin le programme hallucinator[9] requête davantage de bases bibliographiques que bibCheck mais ne distingue pas une référence non trouvée d'une référence hallucinée.

Au milieu de ces nouvelles réalisations, bibCheck apporte plusieurs éléments. En termes d'innovation, bibCheck est capable d'identifier des publications rétractées — en mobilisant le PPS — et de distinguer les références potentiellement hallucinées de celles non trouvées sur Crossref et DataCite. En termes d'implémentation, bibCheck s'utilise sur un site web sans aucune installation ni paramétrage, ce qui facilite son utilisation par le grand public. Un jeu de données de validation a également été constitué pour évaluer les performances de cet algorithme.

Cet article présente donc deux contributions : l'algorithme bibCheck et une collection de test employée pour évaluer la performance de ce dernier.

3. Méthode : l'algorithme bibCheck pour éprouver la qualité des bibliographies

La section 3.1 introduit l'algorithme et la section 3.2 commente ses performances.

3.1. *Validation des références d'une bibliographie soumise à bibCheck*

L'algorithme met en œuvre un ensemble de règles. Tout d'abord, il détecte la présence d'un DOI dans le texte de chaque référence bibliographique, grâce à une expression régulière[10]. Notons que nous prenons en compte des erreurs possibles dues à la retranscription des DOI depuis des PDF. Les deux erreurs les plus courantes

6. https://www.istex.fr/un-autre-web-service-autour-des-references-citees-bibcheck/
7. https://github.com/lidianycs/cerca/tree/v1.1-alpha
8. https://github.com/micha-blip/Simple-article-reference-checker
9. https://github.com/gianlucasb/hallucinator
10. Recommandée par Crossref : https://doi.org/10.64000/cc6d3-tkc85

concernent des DOI coupés par un saut de ligne dans les références, d'une part, et des traits de soulignement transformés en espaces, d'autre part. Puis, trois scénarios sont possibles selon qu'un DOI d est présent dans le texte de la référence r ou pas :

1) d est présent dans r, comme dans la norme APA (figure 1). Il est associé à r' dans Crossref ou DataCite — test effectué via l'API Crossref dédiée ou MetaDoRe, un miroir de DataCite hébergé à l'Inist-CNRS. L'algorithme compare les champs de r et de r'.

2) r ne mentionne pas d, comme dans la norme Vancouver (figure 1). Dans ce cas, les champs de r sont utilisés pour rechercher la référence dans Crossref. C'est une approche utilisée précédemment pour découvrir les articles liés à des preprints d'intérêt (Cabanac *et al.*, 2021b).

3) d est présent dans r, comme dans la norme APA (figure 1). Cependant, d n'est pas présent dans Crossref ou Datacite. Dans ce cas, l'algorithme applique le deuxième scénario pour effectuer une vérification plus rigoureuse de la référence. La présence d'un DOI non attribué ou erroné suggère une probabilité d'hallucination élevée.

Deux possibilités se présentent, selon que la référence r à tester mentionne un doi ou non. Pour ces deux cas, l'algorithme calcule un score de confiance qui dépend des quatre champs bibliographiques d'une référence bibliographique : titre, auteurs, date et support de publication.

Pour chaque champ bibliographique, l'algorithme estime sa similarité vis-à-vis des champs trouvés dans les références candidates restituées par la source interrogée. Pour apparier la référence r à vérifier et les références candidates, une similarité partielle $\mathrm{sim}(a, b) \in [0, 1]$ associée à la distance de Levenshtein est calculée (équation 1) entre les deux chaînes de caractères nettoyées (en minuscule, sans accents et sans caractères non alphanumériques) où :

– $|a|$ est la longueur de la chaîne de caractères a,

– a et b sont deux chaînes de caractères telles que $|a| \geqslant |b|$,

– $\mathrm{dist}(a, b) \in [0, 1]$ est la distance de Levenshtein (Navarro, 2001, p. 37) entre a et b (avec un poids de substitution égal à 2).

$$\mathrm{sim}(a, b) = \max_{t \subseteq a} \left(1 - \frac{\mathrm{dist}(b, t)}{|b| + |t|} \right) \quad [1]$$

Enfin, précisons les conditions nécessaires pour établir un appariement, pour chaque champ bibliographique :

– **titre** : calcul de la similarité partielle sim (équation 1) entre le titre des métadonnées Crossref/MetaDore et la référence bibliographique complète. Le calcul des similarités partielles floues s'opère avec la bibliothèque Python *thefuzz* [11]. Les deux

11. https://github.com/seatgeek/thefuzz

titres correspondent lorsque cette similarité dépasse 0,8. Ce seuil a été fixé sur des exemples et est à ajuster sur un autre corpus que celui de l'évaluation.

– **auteurs** : Si le nom pré-traité, apparaît exactement dans la référence bibliographique complète, le critère est considéré comme validé.

– **date** : l'année est extraite de la date et sa correspondance exacte avec l'année de la référence candidate valide ce critère.

– **support de publication** : calcul de la similarité partielle sim (équation 1) entre la source des métadonnées Crossref/MetaDore et l'ensemble de la référence bibliographique est réalisé. Le critère est validé si cette similarité dépasse 0,8. Parfois la source apparaît de manière abrégée dans la référence (par exemple *Journal of the American Chemical Society* peut être abrégé en *JACS* ou en *J. Am. Chem. Soc.*). Ces abréviations sont gérées par l'algorithme.

Ces 4 critères ne doivent pas nécessairement être validés simultanément, en fonction de la présence et de la validité d'un DOI dans la référence soumise à vérification. Les critères sont nécessairement plus stricts si un DOI est présent mais introuvable. Les différents statuts obtenus sont détaillés dans la section 3.2. Un résumé de l'algorithme est illustré par l'arbre de décision en figure 2.

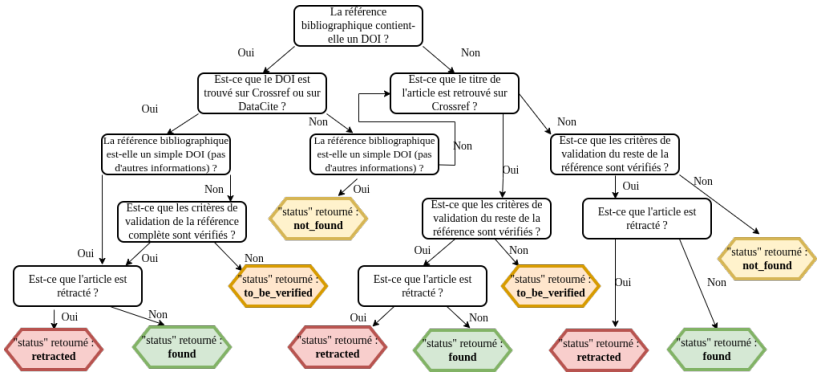


FIGURE 2 : Représentation simplifiée de l'algorithme bibCheck sous forme d'arbre de décision. Étant donnée une référence bibiographique, bibCheck détermine son statut : valide (*found*, en vert), non trouvée (*not_found*, en jaune) à vérifier (*to_be_verified*, en orange) ou rétractée (*retracted*, en rouge).

3.2. *Présentation des différents statuts d'une référence analysée par bibCheck*

Nous mobilisons trois sources pour évaluer le statut d'une référence : Crossref, DataCite et le PPS. Crossref et DataCite sont deux organisations qui attribuent des identifiants pérennes (DOI), respectivement aux publications et aux données de recherche. Certaines publications sont identifiées par des DOI DataCite, elles sont donc présentes

dans DataCite mais ne le sont pas dans Crossref. C'est pourquoi il est pertinent d'interroger ces deux sources de métadonnées bibliographiques. Une fois les références identifiées, nous vérifions qu'elles ne sont pas rétractées. C'est à ce stade que nous utilisons le PPS (Cabanac *et al.*, 2022). Il s'agit d'une base de données collectant des publications suspectes. Nous n'utilisons actuellement que le sous-ensemble du PPS qui liste les DOI des articles rétractés. L'algorithme bibCheck attribue à chaque référence testée un statut parmi les 4 suivants :

– `found`, pour les références trouvées sur Crossref ou DataCite ;

– `retracted`, pour les références rétractées selon le PPS ;

– `to_be_verified`, pour les références soupçonnées d'être générées par IA ;

– `not_found`, pour les références non trouvées. Cette catégorie englobe l'ensemble des références valides qui ne possèdent pas de DOI Crossref ou DataCite. Ces références ne présentent pas de caractéristique d'hallucination.

Nous détaillons désormais les traitements opérés selon que la référence r à tester mentionne un DOI d ou non.

3.2.1. *Un DOI est présent dans la référence et valide*

La recherche du DOI s'effectue d'abord via l'API Crossref. En cas d'échec, une requête est adressée à l'API MetaDore (un miroir de DataCite). Si un DOI est présent et valide, 3 des 4 critères énumérés ci-dessus doivent être remplis, ou alors 2 dont le titre, avec un seuil de similarité supérieur ou égal à 0,7.

Dans ces deux cas, la référence est valide et trouvée, et l'algorithme classe la référence comme `found`. Si ce n'est pas le cas, elle est considérée comme hallucinée avec un DOI valide, et bibCheck la classe comme `to_be_verified`. Si le DOI est mentionné dans le PPS comme rétracté, l'algorithme la classe comme `retracted`.

3.2.2. *Un DOI n'est pas présent dans la référence ou est invalide*

Si aucun DOI n'est présent ou si celui qui est présent n'est pas valide, l'API Crossref est interrogée pour obtenir les références bibliographiques les plus similaires à celles à vérifier. Crossref s'appuie sur Elasticsearch[12] pour trier les résultats par ordre décroissant de pertinence. Pour chacune des cinq premières références, une vérification est effectuée afin de déterminer si l'une d'elles satisfait l'un des 3 cas suivants :

– trois des quatre critères énumérés ci-dessus sont validés (titre, premier auteur, date et support de publication) ;

– deux des quatre critères énumérés ci-dessus sont validés, avec un titre parfaitement identique (la similarité entre les deux titres est supérieure à 0,98). Le titre est donc compté dans ces deux critères à la différence de l'item suivant ;

– seuls deux critères autres que le titre sont validés, avec une similarité entre le titre trouvé et le titre saisi supérieure à 0,6.

12. https://doi.org/10.64000/nxwqn-x9m73

La première référence candidate parmi les cinq répondant à l'un de ces cas est donc retenue. La référence est classée comme `found`, ou `retracted` si le DOI associé à la publication se trouve dans le PPS.

Si un titre provenant de l'une des cinq références correspond parfaitement (similarité supérieure à 0,9 avec la référence saisie) mais qu'aucun autre critère n'est validé, la référence est considérée comme hallucinée. Si la véritable référence ne peut être trouvée parmi les quatre autres, la référence est classée `to_be_verified`. Sinon, la référence n'a pas été trouvée et est classée `not_found`.

La référence est considérée comme générée par l'IA dans tous les cas où aucune des cinq références ne correspond si un DOI invalide a été trouvé dans la référence d'entrée. Dans ce cas, la réponse de l'algorithme est `to_be_verified`. Les erreurs liées aux DOI sont donc pénalisées par notre algorithme.

4. Utilisation, mise en production et accès à bibCheck

L'algorithme bibCheck est utlisable par plusieurs biais. Il peut être exploité sous forme de service web, accessible via une API en transmettant les données au format JSON, sans nécessiter de configuration particulière. Une alternative consiste à l'utiliser via la plateforme dédiée, TDM Factory, qui propose une interface utilisateur simplifiée, adaptée aux non-experts. La figure 3 montre un exemple de résultat obtenu (sur la référence présentée en figure 1 et sur une référence hallucinée) et la figure 4 récapitule la pipeline utilisée.

id	value/doi	value/status	value/reference	value/reference_found
0	10.1038/d41586-021-00733-5	found	Else, H., & Van Noorden, R. (2021). The fight against fake-paper factories that churn out sham science. Nature, 591(7851), 516–519. https://doi.org/10.1038/d41586-021-00733-5	FROM CROSSREF > Else et al. (2021). The fight against fake-paper factories that churn out sham science. Nature. 10.1038/d41586-021-00733-5.
1		to_be_verified	Chen, C. H., & Lu, R. F. (2003). Logistic regression analysis in classification of large datasets. Pattern Recognition Letters, 24(4–5), 511–521. DOI : 10.1016/S0167-8655(02)00207-2	

FIGURE 3 : Un exemple de sortie de l'algorithme bibCheck. Ici la première référence a été trouvée (comme l'indique le statut `found`), et son DOI extrait dans la colonne *value/doi*. La référence analysée est dans la colonne *value/reference* et peut être comparée par l'utilisateur avec la référence trouvée dans une base en ligne (en colonne *value/reference_found*). Elle a été trouvée dans Crossref. La deuxième référence est hallucinée (comme l'indique le statut `to_be_verified`).

L'utilisation de bibCheck sous forme de service web s'explique par la modularité de la plateforme TDM Factory, qui permet d'y intégrer aisément tout algorithme respectant des formats précis pour les entrées et les sorties. Cette approche facilite non seulement la mise en production, mais aussi l'adaptation des outils à différents formats de données.

En utilisant ce procédé, notre algorithme peut s'adapter à différents formats : dans le cas de bibCheck, il est possible de soumettre un article au format PDF ou une liste de références au format CSV. Cela permet en particulier à des utilisateurs d'utiliser

bibCheck sur un export CSV de leur bibliographie Zotero[13]. Le service web traite les références une par une. Cependant, lorsque l'utilisateur fournit un PDF qu'il dépose, ce dernier est analysé par GROBID (2008–2025) pour extraire toute les références. Ensuite l'algorithme effectue le traitement standard référence par référence.

Une fois le fichier à analyser déposé, il est traité sur nos serveurs, de manière anonyme et sécurisée. L'utilisateur peut demander à être notifié par mail à la fin du traitement ou patienter en ligne. Les résultats sont à télécharger sous 7 jours.

En février 2026, à l'heure de la rédaction de cet article, le site TDM Factory est accessible aux agents CNRS ainsi qu'aux ayant droits Istex[14]. Cependant nous rappelons que le code de bibCheck est ouvert et quiconque peut l'exécuter[15].

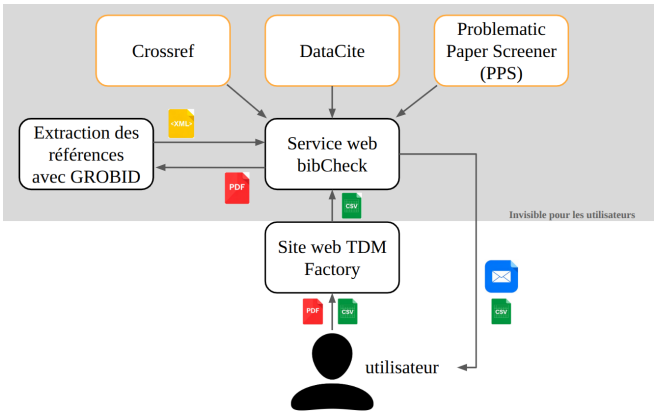


FIGURE 4 : Illustration de l'utilisation de l'algorithme bibCheck via le site web TDM Factory. Les ressources sont encadrées en orange, les traitements non visibles par l'utilisateur sont dans la zone grise. Les icônes représentent les formats des données, et l'icône mail indique que le résultat est un fichier CSV envoyé par mail à l'utilisateur.

5. Évaluation et validation de l'algorithme bibCheck de vérification de références

La section 5.1 présente la composition du corpus d'évaluation ; les résultats sont commentés dans la section 5.2.

5.1. *Constitution du corpus de références bibliographiques en vue de l'évaluation*

Afin d'éprouver bibCheck expérimentalement, nous avons constitué un corpus

13. https://www.zotero.org
14. https://www.istex.fr
15. https://github.com/Inist-CNRS/web-services/tree/main/services/biblio-ref

ouvert[16] de 236 références bibliographiques (Meneghel, 2026), publié sur la plateforme d'outils et de ressources linguistiques Ortolang [17]. Le tableau 1 montre leur répartition selon 3 types suivants :

1) références bibliographiques valides :

- avec un DOI : elles proviennent de Crossref,

- sans DOI : elles proviennent de Crossref mais leur DOI a été supprimé.

2) références bibliographiques hallucinées générées par les LLM : elles ont été soit collectées dans de précédents travaux portant sur les références hallucinées (Alkaissi et McFarlane, 2023; Walters et Wilder, 2023; Chelli *et al.*, 2024), soit manuellement modifiées, à l'aide d'une ou plusieurs de ces opérations :

- modification du DOI ou interversion du DOI avec celui d'une référence d'un sujet proche dans le même domaine,

- ajout d'auteurs probables et/ou suppression d'auteurs associés à l'article,

- modification de la date de parution, modification des numéros de pages concernés ou modification du nom de la revue.

3) références bibliographiques rétractées qui proviennent de Problematic Paper Screener et PubMed ou des maisons d'édition Spandidos, Taylor & Francis et Wiley.

TABLEAU 1 : Répartition par label des références du corpus constitué

Label	Description	avec DOI	sans DOI	Total
found	références valides	51	55	106
to_be_verified	références hallucinées	41	29	70
retracted	références rétractées	60	0	60
Tous	Tout type	152	84	236

5.2. *Résultats de l'évaluation : efficacité de la validation des références*

Pour évaluer bibCheck, nous avons calculé plusieurs métriques. Le tableau 2 montre les précisions, rappels, F-mesures et supports pour chacune des classes. Ces trois métriques sont explicitées dans (Sokolova et Lapalme, 2009) où les noms anglais sont, respectivement, *precision*, *recall* et *F-score*. Le *support* d'une classe correspond au nombre de références évaluées. Nous distinguons les résultats en fonction de la présence ou non d'un DOI dans les données car l'algorithme diffère significativement en fonction de ce dernier. Par exemple, la détection de références hallucinées sans DOI avec des règles est très complexe. Comme attendu, les résultats sont bien meilleurs

16. https://doi.org/10.82270/eval-dataset-bibcheck/v1
17. https://www.ortolang.fr

lorsqu'un DOI est présent dans la référence. En effet, l'algorithme trouve directement la référence associée sur Crossref ou DataCite et une simple comparaison suffit.

Nous constatons que tous les articles rétractés sont prédits correctement comme tels sur ces données de validation. On note aussi qu'aucune donnée labelisée *retracted* sans DOI n'est présente dans le jeu de validation. Il n'est pas nécessaire d'en avoir car en réalité, retrouver une donnée rétractée (selon le PPS) sans DOI équivaut à retrouver une donnée *found* sans DOI (après cela, une simple vérification de présence dans le PPS suffit). Avec ce postulat, on peut estimer la précision à 98 %, le rappel à 75 % et la F-mesure à 85 % pour les références rétractées ne contenant pas de DOI.

TABLEAU 2 : Précision, rappel et F1-score par classe de l'algorithme bibCheck, (a) avec DOI, (b) sans DOI, (c) toutes les références.

(a) Métriques sur l'ensemble des références contenant un DOI

label	Precision	Recall	F1-score	Support
found	0,94	1,00	0,97	51
to_be_verified	1,00	0,90	0,95	41
retracted	1,00	1,00	1,00	60

(b) Métriques sur l'ensemble des références sans DOI

label	Precision	Recall	F1-score	Support
found	0,98	0,75	0,85	55
to_be_verified	0,88	0,24	0,38	29
retracted	0,00	0,00	0,00	0

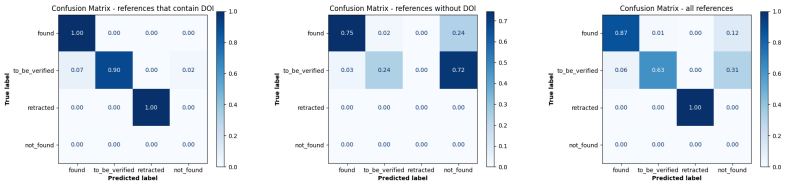
(c) Métriques sur toutes les références

label	Precision	Recall	F1-score	Support
found	0,96	0,87	0,91	106
to_be_verified	0,98	0,63	0,77	70
retracted	1,00	1,00	1,00	60

Quant aux données labelisées *found*, nous rappelons que les données de validation sont toutes disponibles sur Crossref. Sur des données plus variées (par exemple références de conférences sans DOI), on observe beaucoup plus de *not_found* voire de *to_be_verified*.

Regardons à présent le cas des *to_be_verified*. Nous observons un rappel très faible lorsqu'il n'y a pas de DOI, ce qui induit une confusion avec une autre classe. Pour être plus juste et critique dans l'analyse d'erreur, nous avons créé une matrice de confusion, représentée dans la figure 5. Nous la construisons comme suit. Pour chaque donnée, on note y son label et $\hat{y}$ sa prédiction associée par bibCheck (y et $\hat{y}$ peuvent donc prendre 4 valeurs : `found`, `retracted`, `to_be_verified` et `not_found`). Nous calculons pour chaque ligne et colonne (i, j) de la matrice de confusion la probabilité

conditionnelle $P(\hat{y} = j \mid y = i)$. Sur la diagonale, nous calculons les « vrais positifs » et donc nous retrouverons le rappel pour chaque classe. Comme les résultats diffèrent avec et sans DOI, nous distinguons ici aussi ces deux cas.



(a) références contenant un DOI (b) références sans DOI (c) toutes les références

FIGURE 5 : Matrices de confusions de l'algorithme bibCheck, normalisée en ligne, (a) avec DOI, (b) sans DOI, (c) toutes les références. La somme des lignes peut être proche de 1 mais différente pour cause d'arrondis.

Sans surprise, nous avons obtenu un meilleur rappel sur les références *to_be_verified* lorsqu'un DOI est présent (rappel de 0,88 contre 0,19 lorsqu'il n'y a pas de DOI). Mais la matrice de confusion montre que tous les DOI générés par l'IA qui ne sont pas marqués comme *to_be_verified* sont marqués comme *not_found*. La confusion vers cette classe, bien que très grande, est la moins grave : nous partons du principe que ces deux classes seront vérifiées par les utilisateurs. Catégoriser une donnée *to_be_verified* de valide est plus problématique. Le score reste très faible et nous devons l'améliorer tout en restant frugal en termes de ressources de calculs.

Une autre confusion flagrante dans cette matrice est celle des données labelisées *found* avec la classe *not_found* lorsqu'il n'y a pas de DOI (24 % des cas), et une légère confusion avec les *to_be_verified* (2 %).

6. Conclusion

L'algorithme bibCheck répond efficacement à la problématique de la propagation d'articles scientifiques rétractés et de références hallucinées. D'une part, sa simplicité d'utilisation le rend accessible à un grand public. D'autre part, la frugalité de l'algorithme (qui utilise des règles et des appels API) permet un traitement rapide et une utilisation sur un grand volume de références bibliographiques. Les références dotées de DOI sont relativement simples à classifier. En revanche, la détection des hallucinations, notamment pour celles dépourvues de DOI, laisse une marge d'amélioration significative. Il reste essentiel de garantir la scalabilité des solutions proposées pour ces cas. Depuis la version initiale, nous avons corrigé les erreurs de retranscription des DOI par GROBID et intégré DataCite comme source de vérification complémentaire. À court terme, une méthode dédiée aux articles issus d'arXiv est en cours de développement. Nous souhaitons également rendre les résultats explicables pour les utilisateurs, en listant les champs qui semblent différer.

Pour affiner nos résultats, nous souhaitons exploiter d'autres jeux de données afin de comparer nos performances avec celles d'autres travaux. Un travail est initié afin de comparer les résultats d'autres logiciels similaires sur notre corpus d'évaluation pour situer bibCheck vis-à-vis de l'état de l'art. Actuellement, notre focus sur les classes *to_be_verified* et *retracted* rend toute comparaison exhaustive impossible, le seul jeu de données similaire identifié étant insuffisant. Nous prévoyons donc d'enrichir nos données de validation selon deux axes : premièrement, avec des données valides absentes de Crossref, ce qui permettrait d'estimer la confusion de l'algorithme entre la classe *not_found* et les autres. Deuxièmement, étoffer le corpus en y intégrant davantage de références, idéalement issues de sources variées, notamment des références hallucinées. Ce billet d'actualité[18] détaille la façon d'utiliser ce service simplement, via TDM Factory. Il est actuellement en production et utilisé par le CNRS, l'INRAE et l'Académie des sciences par exemple.

Reproductibilité des résultats

Le présent article évalue la version 3.2.0 de bibCheck[19] sur ce corpus [20]. Les codes pour obtenir les résultats sont disponibles sur ce dépôt git[21].

Bibliographie

Aksenfeld R., "Challenge accepted: A reader wrote a program to find fake references in books", 2025. URL https://retractionwatch.com/?p=133193.

Alkaissi H., McFarlane S. I., "Artificial Hallucinations in ChatGPT: Implications in Scientific Writing", *Cureus*, vol. 15, n° 2, p. e35179, February, 2023. DOI: 10.7759/cureus.35179.

Ball P., "Computer conference welcomes gobbledegook paper", *Nature*, vol. 434, n° 7036, p. 946, 2005. DOI: 10.1038/nature03653.

Cabanac G., "Chain retraction: How to stop bad science propagating through the literature", *Nature*, vol. 632, n° 8027, p. 977-979, 2024a. DOI: 10.1038/d41586-024-02747-1.

Cabanac G., "Dépollution de la littérature scientifique : signalement de publications scientifiques non fiables grâce au *Problematic Paper Screener*", *Innovations & Thérapeutiques en Oncologie*, vol. 10, n° 2, p. 124-127, 2024b. DOI: 10.1684/ito.2024.431.

Cabanac G., Clausse A., Jégou L., Maisonobe M., "The geography of retracted papers: Showcasing a Crossref–Dimensions–NETSCITY pipeline for the spatial analysis of bibliographic data", *STI'23: 27th International Conference on Science, Technology and Innovation Indicators*, 2023. https://dapp.orvium.io/deposits/6442fee5c93d17c257de17d2/view.

18. https://www.inist.fr/nos-actualites/tdm-factory-linterface-dediee-a-la-fouille-de-textes
19. https://github.com/Inist-CNRS/web-services/tree/ws-biblio-ref%403.2.0/services/biblio-ref
20. https://hdl.handle.net/11403/eval-dataset-bibcheck/v1
21. https://github.com/Inist-CNRS/ws-data/tree/master/biblio-ref

Cabanac G., Labbé C., “Prevalence of nonsensical algorithmically generated papers in the scientific literature”, *Journal of the Association for Information Science and Technology*, vol. 72, n° 12, p. 1461-1476, 2021. DOI: 10.1002/asi.24495.

Cabanac G., Labbé C., Magazinov A., “Tortured phrases: A dubious writing style emerging in science. Evidence of critical issues affecting established journals”, 2021a. arXiv preprint. DOI: 10.48550/arXiv.2107.06751.

Cabanac G., Labbé C., Magazinov A., “The ’Problematic Paper Screener’ automatically selects suspect publications for post-publication (re)assessment”, 2022. Presented at WCRI 2022: 7th World Conference on Research Integrity. DOI: 10.48550/arXiv.2210.04895.

Cabanac G., Oikonomidi T., Boutron I., “Day-to-day discovery of preprint–publication links”, *Scientometrics*, vol. 126, n° 6, p. 5285-5304, 2021b. DOI: 10.1007/s11192-021-03900-7.

Chelli M., Descamps J., Lavoué V., Trojani C., Azar M., Deckert M., Raynier J.-L., Clowez G., Boileau P., Ruetsch-Chelli C., “Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis”, *Journal of Medical Internet Research*, vol. 26, p. e53164, May, 2024. DOI: 10.2196/53164.

Clausse A., Cabanac G., Cuxac P., Labbé C., “Extraction d’acronymes torturés dans la littérature scientifique”, *Atelier TextMine de la conférence Extraction et Gestion des Connaissances (EGC) de 2024*, Dijon (Bourgogne), France, January, 2024. https://hal.science/hal-04426448.

Cossio M., “A comprehensive taxonomy of hallucinations in Large Language Models”, 2025. arXiv preprint. DOI: 10.48550/arXiv.2508.01781.

Eiko, “Using GPT-3 to search for scientific "references" » Eiko Fried”, January, 2023. https://eiko-fried.com/using-gpt-3-to-search-scientific-references/.

Else H., Van Noorden R., “The fight against fake-paper factories that churn out sham science”, *Nature*, vol. 591, n° 7851, p. 516-519, 2021. DOI: 10.1038/d41586-021-00733-5.

Gaillard L., Bonvallot V., Cuxac P., Parmentier F., “TDM Factory : rendre accessibles des algorithmes de fouilles de textes sans connaissances a priori ni paramétrages”, *Revue des Nouvelles Technologies de l’Information*, vol. Extraction et Gestion des Connaissances, RNTI-E-42, p. 537-544, 2026.

Gravel J., D’Amours-Gravel M., Osmanlliu E., “Learning to fake It: Limited responses and fabricated references provided by ChatGPT for medical questions”, *Mayo Clinic Proceedings: Digital Health*, vol. 1, n° 3, p. 226-234, 2023. DOI: 10.1016/j.mcpdig.2023.05.004.

GROBID, “GeneRation Of BIbliographic Data”, 2008–2025. https://github.com/kermitt2/grobid.

Guenci M., Heibi I., Parravicini C., Peroni S., Soricetti M., “A pipeline for matching bibliographic references with incomplete metadata: experiments with Crossref and OpenCitations”, 2025. arXiv preprint. DOI: 10.48550/arXiv.2511.18408.

Labbé C., Cabanac G., West R. A., Gautier T., Favier B., Byrne J. A., “Flagging incorrect nucleotide sequence reagents in biomedical papers: To what extent does the leading publication format impede automatic error detection?”, *Scientometrics*, vol. 124, n° 2, p. 1139-1156, August, 2020. DOI: 10.1007/s11192-020-03463-z.

Labbé C., Labbé D., Portet F., “Detection of Computer-Generated Papers in Scientific Literature”, *in* M. Degli Esposti, E. G. Altmann, F. Pachet (eds), *Creativity and Universality in Language*, Lecture Notes in Morphogenesis, Springer, p. 123-141, 2016. DOI: 10.1007/978-3-319-24403-7_8.

Lammey R., Oransky I., “Retraction Watch and Crossref: Collaborating to Improve the Assessment of Scholarly Outputs”, *Science Editors*, 2024. DOI: 10.36591/se-d-4701-01.

Liang W., Zhang Y., Wu Z., Lepp H., Ji W., Zhao X., Cao H., Liu S., He S., Huang Z., Yang D., Potts C., Manning C. D., Zou J., “Quantifying large language model usage in scientific papers”, *Nature Human Behaviour*, vol. 9, nº 12, p. 2599-2609, 2025. DOI: 10.1038/s41562-025-02273-8.

Liu Q., Barhoumi A., Labbé C., “Miscitations in scientific papers: dataset and detection”, *Proceedings of the 14th Bibliometric-enhanced Information Retrieval workshop (BIR@ECIR'24)*, CEUR, p. 53-65, 2024. URL: https://ceur-ws.org/Vol-3989/paper-05.pdf.

Meneghel V., “TDM Evaluation Dataset - bibCheck”, 2026. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr. DOI: 10.82270/eval-dataset-bibcheck.

Naddaf M., Quill E., “Hallucinated citations are polluting the scientific literature. What can be done?”, *Nature*, vol. 652, nº 8108, p. 26-29, 2026. DOI: 10.1038/d41586-026-00969-z.

Navarro G., “A guided tour to approximate string matching”, *ACM Computing Surveys*, vol. 33, nº 1, p. 31-88, 2001. DOI: 10.1145/375360.375365.

O'Grady C., “Software that detects ‘tortured acronyms’ in research papers could help root out misconduct”, *Science*, vol. 384, nº 6700, p. 1056, 2024. DOI: 10.1126/science.znqe1aq.

Sakai Y., Kamigaito H., Watanabe T., “HalluCitation matters: Revealing the impact of hallucinated references with 300 hallucinated papers in ACL conferences”, January, 2026. arXiv preprint. DOI: 10.48550/arXiv.2601.18724.

Sokolova M., Lapalme G., “A systematic analysis of performance measures for classification tasks”, *Information Processing & Management*, vol. 45, nº 4, p. 427-437, 2009. DOI: 10.1016/j.ipm.2009.03.002.

Tramèr F., “Trends in LLM-Generated Citations on arXiv”, August, 2025. https://spylab.ai/blog/hallucinations/.

Van Noorden R., “More than 10,000 research papers were retracted in 2023 — a new record [News]”, *Nature*, vol. 624, nº 7992, p. 479-481, 2023. DOI: 10.1038/d41586-023-03974-8.

Van Noorden R., “These universities have the most retracted scientific articles [News]”, *Nature*, vol. 638, nº 8051, p. 596-599, 2025. DOI: 10.1038/d41586-025-00455-y.

Walters W., Wilder E., “Fabrication and errors in the bibliographic citations generated by ChatGPT”, *Scientific Reports*, vol. 13, p. article 14045, 2023. DOI: 10.1038/s41598-023-41032-5.

Exploitation des signaux faibles dans les conversations familiales comme levier pour enrichir les données structurées en contexte d'aidance

Paul Cariou[1,2], **Kaoutar Akhsass**[2,3], **Luiz Angelo Steffenel**[3], **Manuele Kirsch Pinheiro**[2]

1. Tamalou Family
75005, Paris, France
Paul@tamalou.io

2. Centre de Recherche en Informatique
Université Paris 1 Panthéon Sorbonne, Paris France
Kaoutar.akhsass@univ-paris1.fr, Manuele.kirsch-pinheiro@univ-paris1.fr

3. LICIIS/LRC CEA DIGIT, Université de Reims Champagne-Ardenne
Reims, France
Luiz-angelo.steffenel@univ-reims.fr

RESUME. Les dossiers médicaux et comptes rendus de consultation négligent souvent les signaux faibles de la vie quotidienne (sommeil, humeur, ressentis, effets secondaires des traitements), se concentrant davantage sur les marqueurs biologiques ou les événements aigus. Pourtant, ces données « invisibles », essentielles pour ajuster les soins, sont souvent partagées de manière informelle entre les proches. Ces échanges informels représentent une source significative de données et peuvent être retrouvés sur des applications dédiées, comme celle proposée par l'entreprise Tamalou, qui offre une application gratuite de coordination des soins conçue pour réduire la charge mentale de l'aidant principal, notamment dans le cadre des soins à domicile. Cela soulève la question de savoir comment les nouveaux outils d'IA, et en particulier les grands modèles de langage (LLM), peuvent structurer ces données et les rendre exploitables par les équipes soignantes. Ces modèles pré-entraînés peuvent être utilisés pour contextualiser et évaluer les symptômes et ressentis exprimés de manière explicite ou implicite dans un langage informel. Dans cet article, nous proposons d'explorer l'utilisation d'un LLM pour identifier les signaux faibles cachés dans ces échanges. Nous présentons ici nos premiers résultats, obtenus à partir d'un corpus d'environ 2 000 verbatims et évalués grâce à une approche de type LLM-as-a-Judge. À terme, cette proposition s'intégrera à l'application Tamalou, ouvrant la possibilité d'ajouter les données obtenues aux informations du patient, notamment par la génération de notes structurées compatibles avec le Dossier Médical Partagé (DMP).

MOTS-CLES : Large Language Models (LLM), IA médicale, structuration de données, coordination familiale, données de santé, Natural Language Processing (NLP), signaux faibles.

1. Introduction

Le plus souvent, les dossiers médicaux se concentrent sur des données médicales formelles et standardisées, telles que les marqueurs biologiques et les événements aigus. Ils éludent souvent les signaux faibles de la vie quotidienne. Nous entendons par signaux faibles les informations informelles et subjectives partagées dans les échanges familiaux, qu'elles soient d'ordre physique (*e.g.* troubles du sommeil, effets secondaires des traitements) ou émotionnel (*e.g.* découragement, anxiété). Ces données « invisibles » peuvent s'avérer particulièrement pertinentes pour ajuster les soins. En effet, leur exploitation pourrait permettre aux équipes soignantes d'identifier précocement des complications (*e.g.* dépression, effets secondaires non déclarés) ou d'ajuster les traitements en fonction de l'évolution des symptômes et de l'état émotionnel du patient.

La disponibilité des échanges informels entre proches sur des plateformes numériques, telles que Reddit, les forums de discussion santé ou des applications dédiées comme Tamalou[1], ouvre de nouvelles perspectives pour les soins. Grâce aux avancées récentes en Traitement Automatique du Langage Naturel (TALN), et notamment aux grands modèles de langage (*Large Language Models* _ LLM), il est désormais envisageable d'identifier automatiquement des signaux faibles particulièrement pertinents pour les équipes soignantes.

Cet article explore le potentiel du TALN et des LLM pour structurer ces données informelles et les rendre exploitables par les professionnels de santé. La contribution principale réside dans la démonstration de la faisabilité d'un pipeline basé sur les LLM, capable d'extraire des signaux faibles à partir de conversations informelles en utilisant des techniques prêtes à l'emploi. Nous proposons d'extraire des symptômes à partir de messages courts échangés sur des forums de santé publics. L'analyse de ce type de messages présente plusieurs défis : leur concision n'offre peu voire pas de contexte pour l'analyse ; la vocation généraliste des forums ne donne pas de focus sur une pathologie spécifique ; les auteurs sont peu identifiables, et les descriptions sont souvent effectuées par des patients ou proches sans connaissance médicale. Si à terme ce travail vise à intégrer ces données dans des outils cliniques comme le Dossier Médical Partagé (DMP), cet article se concentre sur la validation technique et méthodologique de l'extraction des données.

Plus largement, notre travail s'inscrit dans un cadre méthodologique d'*Action Design Research* (ADR) (Sein et al. 2011), qui lie la conception d'un artefact technique à la production de connaissances théoriques en Systèmes d'Information. L'artefact développé ici est un pipeline de traitement automatique du langage naturel (TALN) dédié à l'extraction de signaux faibles. Le problème traité concerne l'absence de capture des données informelles du quotidien par les outils de santé structurés (Greenhalgh et al., 2009), dans un contexte organisationnel centré sur la plateforme Tamalou et ses groupes de proches aidants. Les théories mobilisées incluent la théorie de la charge des aidants (Montgomery et al., 1985 ; Zarit et al., 1980), la dynamique

[1] https://www.tamalou.io

des émotions en tant que processus temporel (Kuppens et Verduyn, 2017), ainsi que la conscience de groupe dans les collecticiels (Dourish et Belotti, 1992). La contribution de cet article se situe à l'intersection de l'amélioration de l'extraction des données, par l'application des LLM au contexte spécifique et peu exploré des aidants familiaux francophones, et de l'innovation, avec un pipeline d'extraction inédit fondé sur des construits cliniques validées.

Le reste de l'article est structuré comme suit : la Section 2 aborde la littérature et le contexte du travail, tandis que la Section 3 décrit la méthodologie proposée. La Section 4 présente les résultats obtenus et la Section 5 discute ces résultats, leurs limites et enjeux éthiques. Enfin, la Section 6 conclut et propose des pistes pour de recherches futures.

2. Contexte du travail

2.1. Coordination familiale dans un contexte d'aidance

Dans un contexte d'interactions dématérialisées, les dynamiques de groupe numérique deviennent essentielles. Les groupes familiaux sont confrontés au défi de maintenir une conscience collective, ou *group awareness* (Dourish et Belotti, 1992), définie comme la capacité des membres à percevoir et comprendre les activités, intentions et états émotionnels des autres participants. Ce concept, central aux collecticiels (Grudin, 1994 ; Ellis et al. 1991), permet d'assurer une coordination efficace et d'éviter les doublons dans les tâches. Dans le cadre des soins à domicile, Klasnja et Pratt (2012) ont montré que les technologies mobiles de santé peuvent soutenir la coordination familiale. Cette coordination est d'autant plus importante que des études récentes (Smriti et al., 2024) révèlent qu'une majorité des aidants souffrent d'une détresse psychologique significative, aggravée par la charge émotionnelle et la complexité des dynamiques relationnelles intrafamiliales. Ces résultats font écho aux travaux fondateurs de Montgomery et al. (1985), qui ont formalisé la distinction entre charge objective et subjective, posant ainsi les bases d'une mesure systématique de la charge des aidants. Néanmoins, l'adoption d'outils collaboratifs dépend fortement de leur adaptation aux pratiques informelles des groupes, afin de pallier la tension entre formalisation technologique et pratiques informelles. Il est donc essentiel que ces outils puissent proposer une réelle plus-value aux familles et aux soins.

2.2. Modèles de TALN cliniques pré-entrainés

Les nouvelles techniques liées aux LLM offrent une opportunité d'exploiter les signaux faibles présents dans les messages et autres documents. La littérature montre des avancées significatives, depuis les tâches d'extraction comme le challenge i2b2 (Sun et al., 2013), jusqu'aux jeux de données annotés par des médecins, tels que MedNLI (Romanov et al., 2018), en passant par les modèles pré-entraînés sur des textes cliniques, comme Clinical BERT (Alsentzer et al., 2023), BERT LARGE (Si et al. 2019), SCAI-BIO/BioGottBERT (Diaz Ochoa et al., 2025), et Clinical

ModernBERT (Lee et al., 2025). Ces modèles promettent des performances améliorées dans l'extraction d'entités cliniques et la détection de symptômes. Cependant, les textes utilisés pour l'ajustement fin (*fine-tuning*) de ces modèles sont souvent rédigés par des professionnels de santé, employant un vocabulaire précis et des descriptions moins ambiguës que celles des messages échangés entre les proches de patients. Par ailleurs, lorsqu'on considère les observations des proches, un autre défi réside dans la collecte des données : bien que des plateformes supportant ces échanges existent, notamment des collecticiels (Grudin, 1994 ; Ellis et al. 1991), peu de travaux ciblent spécifiquement les unités familiales et leurs interactions (Branco et al., 2016 ; Mejia et al., 2007 ; Smriti et al., 2024 ; Tang et al., 2018).

2.3. Choix des LLM et apprentissage par prompt

Comme mentionné précédemment, la littérature sur les LLM comprend plusieurs modèles entraînés sur des termes et matériaux cliniques, tels que Clinical BERT (Alsentzer et al., 2019), BERT LARGE (Si et al., 2019), SCAI-BIO/BioGottBERT (Diaz Ochoa et al., 2025), et Clinical ModernBERT (Lee et al., 2025). La plupart de ces modèles partagent un dénominateur commun : ce sont des modèles basés sur BERT, mieux adaptés à des tâches spécifiques telles que la classification, la traduction ou le résumé automatique. Les modèles basés sur GPT, en revanche, ont rapidement évolué ces dernières années, apportant flexibilité et outils itératifs permettant la construction d'applications efficaces sans entraînement spécifique. En effet, les modèles modernes supportent des fenêtres de prompt et de contextes plus larges permettant un raisonnement spécifique sans nécessiter de *fine-tuning*.

Face à la multiplicité des LLM disponibles, GPT4 (Open AI), Gemini (Google), LLaMa (Meta), Mistral (Mistral AI), le choix du modèle doit être motivé par des critères techniques, linguistiques et éthiques. Sur le plan technique, les travaux de Liang et al. (2022) ont évalué les grandes familles de modèles sur des dimensions multiples (précision, robustesse, équité, efficacité) et soulignent qu'aucun modèle ne domine sur tous les axes : les compromis entre capacité générative, contrôlabilité et coût d'inférence sont déterminants selon le contexte applicatif. Dans le domaine médical, Brown et al. (2020) ont évalué GPT-4 sur des tâches ouvertes et des contextes longs, sans recourir au fine-tuning. Cette distinction est déterminante pour notre cas d'usage : l'extraction de signaux cliniques à partir de verbatims informels et hétérogènes requiert une capacité de raisonnement contextuel et de généralisation qu'offrent davantage les architectures des GPTs.

L'utilisation d'un LLM génératif comme Mistral ou GPT-4 nécessite une conception soignée du prompt soumis au modèle. Un prompt mal formulé ou trop ambigu peut conduire à des sorties incohérentes ou hallucinatoires (Ji et al., 2023). La littérature distingue plusieurs paradigmes de *prompting*. L'apprentissage avec contexte (*in-context learning* ou ICL), formalisé par Brown et al. (2020) avec GPT-3, désigne la capacité d'un LLM à résoudre une tâche en se basant uniquement sur des exemples ou des instructions fournis dans le prompt, sans mise à jour des poids du modèle. Cet apprentissage se décline en trois modalités principales : le *zero-shot* (instruction seule, sans exemple), le *one-shot* (un exemple) et le *few-shot* (quelques

exemples). Brown et al. (2020) montrent que les performances des LLM augmentent significativement avec le nombre d'exemples fournis, jusqu'à un certain seuil au-delà duquel le gain marginal diminue.

Au-delà du simple ICL, la littérature propose différentes stratégies de *prompting.* Par exemple, le *Chain-of-Thought prompting* (CoT) (Wei et al., 2022) consiste à guider le modèle pour qu'il détaille les étapes intermédiaires de son raisonnement avant de produire une réponse finale, alors que l'apprentissage par prompt (*prompt learning* ou *prompt tuning*) désigne une approche d'optimisation paramétrique dans laquelle de courts vecteurs continus appelés *soft prompts* sont assimilés par gradient tout en maintenant les poids du LLM figés (Lester et al., 2021). Dans le domaine médical, Lievin et al. (2023) ont appliqué le CoT pour guider GPT-4 sur des questions cliniques ouvertes, observant des gains notables en cohérence et précision. Même si ces techniques permettent des gains plus ou moins importants selon le contexte d'application, elles ont souvent l'inconvénient d'être itératives, demandant une séquence ordonnée de *prompts*.

2.4. LLM-as-a-Judge

Évaluer les modèles de langage et notamment les LLM est un défi majeur, car les benchmarks traditionnels (tels que MMLU ou HELM) ne capturent pas toujours les préférences humaines, notamment pour des tâches ouvertes et conversationnelles. Zheng et al. (2023) proposent une méthode innovante : le LLM-en-tant-que-juge (*LLM-as-a-Judge*), dans laquelle un LLM avancé (tel que GPT-4) évalue les réponses d'autres modèles. Ainsi, au lieu d'un étiquetage manuel ou d'une validation des résultats, la sortie du LLM est soumise à un autre LLM avec un prompt d'évaluation qui note la réponse selon des critères définis par l'utilisateur. L'étude de Zheng et al. (2023) montre que : (i) GPT-4 atteint plus de 80 % d'accord avec les préférences humaines, un niveau comparable à l'accord interhumain ; (ii) les biais (position, verbosité, auto-promotion) peuvent être atténués par des techniques telles que le changement de position ou l'utilisation de références guidées ; et (iii) cette méthode est évolutive, explicable, et moins coûteuse que les évaluations humaines manuelles.

Plusieurs travaux ultérieurs ont approfondi et nuancé ce paradigme. Gu et al. (2022) ont mis en évidence des biais systématiques dans les jugements LLM, notamment le biais de position (préférence pour la première réponse présentée), le biais de verbosité (tendance à favoriser les réponses plus longues) et le biais d'auto-promotion (le modèle tend à préférer ses propres sorties). Ces biais peuvent être partiellement atténués par des techniques dédiées : l'inversion de l'ordre de présentation des réponses, le recours à des références guidées ou l'utilisation de rubriques explicites détaillant les critères d'évaluation (Saha et al., 2023).

Dans le domaine médical, l'approche *LLM-as-a-Judge* présente un intérêt certain, car l'évaluation manuelle par des experts cliniques est à la fois coûteuse et difficile à échelonner. Singhal et al. (2023) utilisent un médecin LLM juge pour évaluer la cohérence clinique, la sécurité et la complétude des réponses médicales générées par leur modèle Med-PaLM 2, observant un accord élevé avec les évaluateurs humains

experts. Nori et al. (2023) ont appliqué un protocole d'évaluation similaire à GPT-4 sur des cas cliniques complexes, confirmant la validité du LLM-juge dans ce contexte. Ces travaux soulignent cependant un point de vigilance : la fiabilité du juge dépend fortement de la qualité du prompt d'évaluation et de la précision des critères fournis.

Enfin, la question des hallucinations, définies comme la génération d'informations factuellement fausses ou non ancrées dans le contexte fourni (Ji et al., 2023), est centrale dans l'usage d'un *LLM-as-a-Judge* appliqué à l'extraction clinique. Gu et al. (2026) énoncent par ailleurs plusieurs recommandations, qui insistent sur la nécessité de mesures dédiées à la fidélité factuelle dans des contextes à fort enjeu.

3. Méthodologie : Identification des Signaux Faibles

Afin de capturer les signaux faibles issues des échanges des membres de famille des patients sur les plateformes collaboratives telles que Tamalou, nous proposons une méthodologie structurée en quatre étapes (voir la Figure 1) : Collecte des données, préparation des données, analyse des verbatims ainsi que l'évaluation des résultats via l'approche *LLM-as-a-Judge*. Cet article a pour objectif d'évaluer ce pipeline qui sert à l'extraction et l'identification des informations cliniques et médicales des patients à partir des échanges de certains aidants. Deux objectifs majeurs ont été précisés : (i) identifier ou extraire les symptômes et maladies mentionnées concernant le patient, accompagné de fragment de texte comme preuve ; et (ii) évaluer la qualité des extractions en utilisant un *LLM-as-a-Judge* qui examine l'encrage dans les preuves, la cohérence du *scoring*, et les hallucinations.

Plusieurs modèles de langage LLM sont aujourd'hui disponibles. Dans ce travail, nous avons donc décidé de nous appuyer sur des modèles bien connus et de développer des prompts et contextes spécifiques pour répondre à nos objectifs. Après avoir testé différents LLM, nous avons opté pour la famille de modèles Mistral, principalement pour des raisons linguistiques et culturelles : ce modèle est conçu et optimisé pour le français, ce qui lui permet de saisir les nuances spécifiques à la langue française et à son contexte social, économique et culturel. Nous estimons qu'il apprécie plus précisément les expressions familières et le phrasé informel utilisés dans les conversations décontractées des conversations du quotidien. De plus, ce LLM est développé par une entreprise européenne (Mistral AI), ce qui offre de meilleures garanties concernant la conformité aux normes RGPD et s'aligne avec notre engagement en faveur de la souveraineté numérique.

Une fois la famille de modèles choisie, la conception des prompts a fait l'objet d'un processus itératif rigoureux où, à chaque étape, les résultats obtenus étaient analysés et le prompt était ajusté en conséquence, permettant ainsi une amélioration progressive et continue des performances du modèle. Par ailleurs, dans notre contexte, les données annotées sont limitées et nous avons privilégié une approche de *prompting* discret (*zero-shot*) afin de préserver la généralité du modèle, assortie d'un schéma de sortie strict, ce qui constitue une forme de contrôle structurel du comportement du modèle sans modifier ses paramètres.

Enfin, concernant notre *LLM-as-a-Judge*, celui-ci a été explicitement instruit de détecter les exemples hallucinés (notés WRONG) et les omissions (éléments présents dans le verbatim mais non extraits), garantissant ainsi une évaluation couvrant à la fois la précision et le rappel de l'extraction. Cette double évaluation s'inscrit dans les recommandations formulées par Gu et al. (2026).

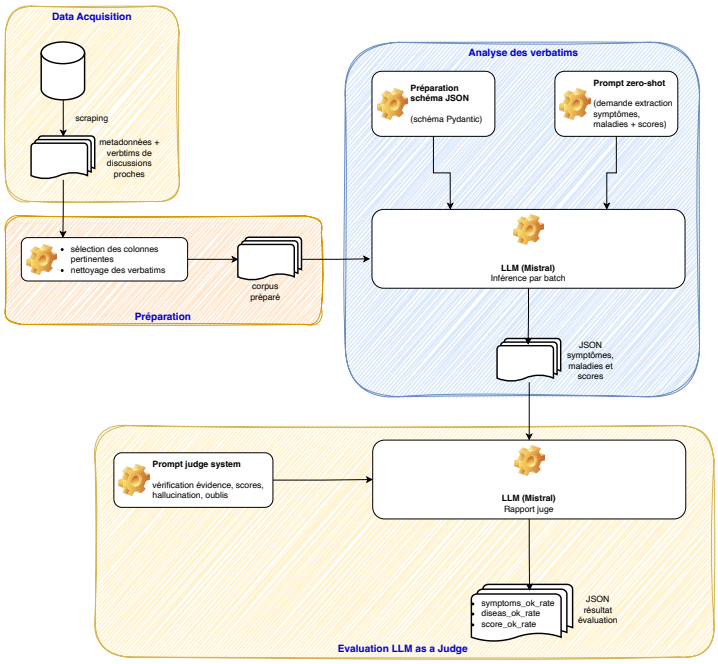


Figure 1. Pipeline d'identification des signaux faibles à partir des échanges informels

3.1. Corpus et considérations éthiques

Dans ce travail, une étape de préparation des données a été réalisée. Nous avons collecté un corpus de 2203 verbatims issus des échanges sur des forums de santé entre proches de personnes rendues vulnérable par l'âge ou une maladie. Ces données, partagées informellement entre proches, couvrent des observations quotidiennes (sommeil, humeur, effets secondaires des traitements) et des expériences émotionnelles (découragement, anxiété, fatigue). Chaque verbatim a été anonymisé et prétraité afin de supprimer les informations personnelles, tout en gardant les indicateurs cliniques et émotionnels pertinents et importants.

Pour assurer la diversité et la représentativité des données, nous avons :

- Sélectionné des verbatims issus de différents contexte familiaux (par exemple, des soignants des personnes âgées, des patients en période de rémission, des familles avec des maladies chroniques) ;
- Équilibré le corpus entre observations physiques (par exemple, « Il a de la fièvre depuis hier ») et signaux émotionnels (par exemple, « Elle ne veut plus voir personne depuis sa chimiothérapie ») ;
- Inclus des cas aberrants (par exemple, verbatims ambigus ou contradictoires) pour tester la robustesse du modèle.

Il est donc important de souligner que le corpus utilisé dans cette étude a été constitué à partir de forums publics sur la santé. Tous les verbatims ont été intégralement anonymisés avant traitement. Aucune autorisation éthique institutionnelle n'était requise pour cette phase au regard de la réglementation française applicable et notamment RGPD, Art. 89.

3.2. Préparation des données : prétraitement et sélection des colonnes

Avant d'inférer, une étape de préparation des données a été réalisé. Au cours de cette étape, nous avons nettoyé et normalisé le corpus afin de réduire le bruit et d'améliorer la robustesse du modèle. Le pipeline de prétraitement a consisté en plusieurs étapes : normalisation des espaces, suppression des observations dupliquées, filtrage des verbatims vides et standardisation de l'encodage. À partir de ce corpus nettoyé, nous avons sélectionné les colonnes ou les variables les plus pertinentes pour cette première analyse et par la suite s'en servir pour construire le LLM *input payload* : {*row_id, date, caregiver, disease, verbatim*}. Ces *payloads* ont été exportés en format JSON pour le traitement par batch ou lots.

3.3. Analyse des verbatims

Afin de procéder à l'analyse des verbatims, un schéma JSON spécialement conçu a été utilisé pour accommoder la sortie du LLM. Afin de formaliser ce schéma, nous avons utilisé Pydantic, une bibliothèque de validation de données pour Python. Ainsi, la tâche d'extraction produit un objet JSON structuré pour chaque ligne d'entrée contenant les champs suivants : *patient_mentioned* (booléen), *patient_symptoms* (liste des symptômes extraits) et *patient_diseases* (liste des maladies extraites). Le système d'extraction cible exclusivement les informations qui se rapportent explicitement au patient. Les émotions, conseils ou descriptions de l'aidant non liées au patient sont exclus de cette extraction. Lorsque le patient n'est pas clairement identifiable ou décrit dans les verbatims, la sortie doit indiquer que le patient n'est pas mentionné, avec les listes de symptômes et de maladies laissées vides.

Afin d'assurer la cohérence structurelle et la validité sémantique, ces variables ont été encodées selon un schéma de *scoring* ordinal contraint, présenté dans le Tableau 1, et selon les règles suivantes, appliquées dans le prompt :

- Dans *patient_symptoms, score* $\in \{1,2\}$ uniquement ;
- Dans *patient_diseases, score* $\in \{3,4\}$ uniquement ;
- Si rien n'est extrait, les listes restent vides et le score est 0.

Tableau 1. Schéma de scoring selon une échelle de preuve à cinq niveaux

Niveau	Description	Exemple
0	Aucune indication clinique ou émotionnelle exploitable	"Elle va mieux aujourd'hui."
1	Symptômes ou ressentis non spécifiques	"J'ai mal au ventre."
2	Symptômes spécifiques et/ou multiples, ou effets secondaires identifiés	"J'ai mal au ventre, des courbatures et de la fièvre."
3	Famille de maladie nommée (sans diagnostic précis)	"Il a un cancer."
4	Maladie ou diagnostic spécifique mentionné	"Elle a la maladie de Crohn."

3.4. Analyse des verbatims : inférence LLM avec Mistral

Nous avons effectué une extraction zéro-shot en utilisant le LLM Mistral (*modèle ministral-8b:latest*) avec un décodage déterministe (*température=0*) pour assurer la reproductibilité. Au lieu d'un *fine-tuning* supervisé, nous avons utilisé l'apprentissage par prompt (apprentissage en contexte zéro-shot), dans lequel le modèle est guidé par des instructions explicites et un format de sortie strictement défini. Le prompt d'extraction spécifiait les points suivants : (i) les entités cibles (symptômes et maladies du patient) ; (ii) des critères d'inclusion stricts (extraction centrée sur le patient avec gestion de la négation) ; et (iii) un schéma JSON structuré (mentionné précédemment) requérant des extraits de preuve pour chaque élément extrait.

Afin de réduire la génération non contrôlée et d'améliorer la traçabilité, nous avons conçu une stratégie de *prompting* multicouche imposant : (1) l'ancrage dans les preuves via des citations textuelles requises (maximum 20 mots) liées aux identifiants de ligne ; (2) la gestion explicite de la négation (*negated = true* le cas échéant) ; (3) une extraction d'attributs structurée ; (4) des plages de *scoring* restreintes ; et (5) une couverture complète en retournant un objet de sortie par *row_id* d'entrée, y compris les sorties vides quand aucune information sur le patient n'est présente. Ces contraintes, combinées à l'échantillonnage déterministe et à la validation du schéma, constituent une spécification formelle légère qui limite le comportement du modèle et réduit le risque d'hallucination. En outre, les lignes d'entrée étaient traitées par lot sous un budget de *tokens* fixe pour maximiser le débit ; lorsqu'une réponse était tronquée, le lot était automatiquement divisé et relancé. Enfin, l'ensemble du pipeline a été implémenté et exécuté sur un nœud NVIDIA DGX Spark (processeur NVIDIA GB10, 128 Go de RAM, 4 To de stockage) à l'Université de Reims Champagne-Ardenne.

4. Évaluation automatisée par le LLM

Afin d'évaluer la qualité de l'extraction structurée produite par Mistral LLM, nous avons mis en place une étape d'évaluation *LLM-as-a-Judge*. Le modèle LLM utilisé pour le juge est un *mistral-large-latest*, également exécuté sur notre nœud DGX Spark. Le juge a reçu, pour chaque échantillon, le verbatim nettoyé de l'aidant et l'extraction structurée proposée (symptômes, maladies, scores et preuves). Il a été instruit, via un prompt dédié, à vérifier que chaque élément extrait est soutenu par des preuves présentes dans le verbatim et à vérifier la cohérence des contraintes de score (cf. Tableau 1). En outre, il a été instruit de détecter les hallucinations (éléments inventés ou non supportés) et les omissions, c'est-à-dire les éléments présents dans le verbatim, mais manquants dans l'extraction.

Enfin, le juge retourne un rapport JSON conforme, lui aussi, à un schéma prédéfini, incluant des étiquettes de correction au niveau de la ligne (OK, PARTIEL ou INCORRECT), des comptages et listes d'éléments hallucinés, d'éléments manquants, et des notes de diagnostic optionnelles.

4.1. Résultats et validation

Nous avons appliqué le *LLM-as-a-Judge* aux observations dans lesquels le modèle d'extraction a estimé que le patient est explicitement mentionné, c'est-à-dire les entrées directement liées au patient. Nous nous sommes concentrés uniquement sur ces verbatims afin d'éviter les erreurs et les mauvaises interprétations de messages qui ne se réfèrent pas nécessairement à un patient. En effet, notre objectif étant d'extraire les symptômes du patient à partir des verbatims, ceux qui font directement référence à un patient sont les plus susceptibles de contenir des symptômes le concernant. Dans notre jeu de données, les messages liés aux patients sont identifiés dans 1 639 entrées (contre 539 entrées considérées comme des messages non liés au patient). Le juge évalue donc la qualité de l'extraction sur ce sous-ensemble où des informations liées au patient sont attendues et extractibles.

À partir de ces sorties, nous avons calculé trois métriques d'agrégation qui fournissent un proxy automatisé pour l'ancrage factuel et la couverture des informations cliniques extraites :

- OK pour les symptômes, maladies et contraintes de score ;
- Hallucinations moyennes par ligne ;
- La moyenne des éléments manquants par ligne.

Lors de l'analyse des messages, la précision de détection est fortement liée aux éléments factuels et observables exprimés dans le texte. Par exemple, les symptômes sont les éléments les plus couramment exprimés, et leur identification est souvent aisée, avec moins de 10 % d'inférences incorrectes (et presque 70 % d'exactitude forte), comme l'illustre la Figure 2(a). Relier les messages aux maladies est quelque peu plus difficile, comme l'exprime la Figure 2(b). Néanmoins, 50,7 % des inférences ont été jugées correctes, 36,2 % partiellement correctes et seulement 13,1 %

incorrectes. Enfin, en ce qui concerne les contraintes de score délimitées dans le Tableau 1, on constate que les proportions de correspondances partielles ou incorrectes augmentent (voir Figure 2(c)). Cela peut être lié aux scores fixes, qui établissent des frontières trop restrictives pour classer les messages textuels (à partir de quand un ensemble de symptômes devient-il l'expression d'une maladie ?).

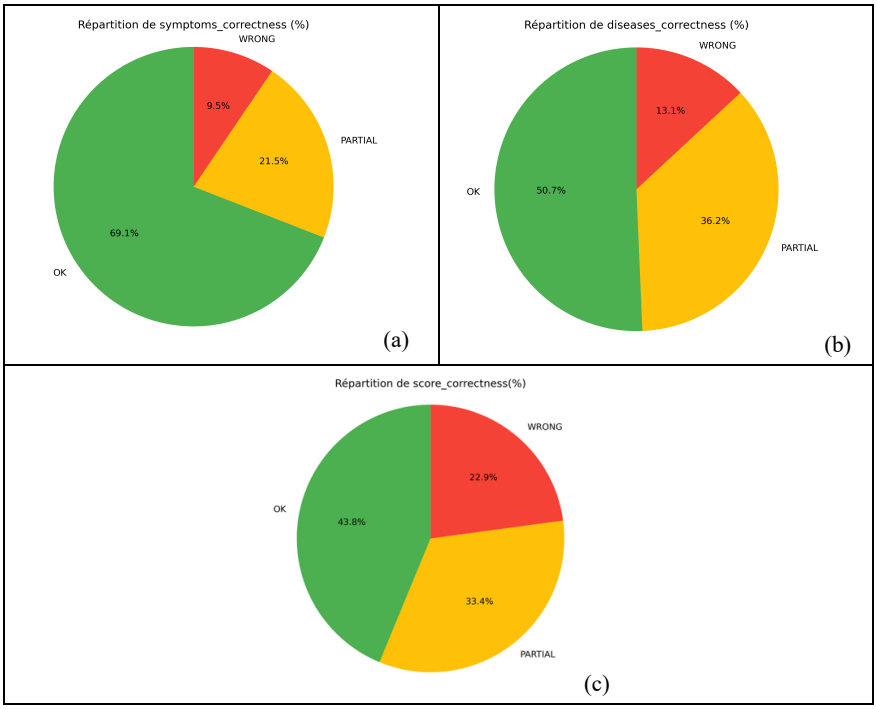


Figure 2. Repartition de la correction des extractions: (a) symptoms, (b) maladies, (c) scores

Une autre métrique utilisée dans cet article est la moyenne des hallucinations par ligne, qui indique combien d'hallucinations (éléments inventés ou non supportés) ont été détectées, en moyenne. Dans le sous-ensemble évalué, ce taux a atteint 0,695, ce qui signifie que moins d'un mot dans un message semble être le résultat d'une hallucination. Une analyse plus approfondie des sorties devra déterminer si les hallucinations sont dominées par des erreurs bénignes (par ex. confusion dans la description d'un symptôme) ou par des erreurs plus graves.

Enfin, la dernière métrique considérée dans ce travail est la moyenne des omissions par ligne, qui indique combien d'éléments des messages originaux n'ont pas été extraits par le LLM. Selon le *LLM-as-a-Judge*, en moyenne seulement 1,39 éléments sont manquants, une faible fraction si l'on considère la longueur moyenne des entrées (en moyenne, plus de 700 mots par verbatim).

5. Discussion

5.1. Interprétation des résultats

Plusieurs biais peuvent être identifiés dans cette expérience initiale. Premièrement, les verbatims proviennent de conversations entre aidants sur des forums d'entraide, et non au sein d'un même cercle familial. L'analyse a donc été menée sur des verbatims isolés, décontextualisés et sans données temporelles longitudinales. Il s'agit là d'une limite inhérente à la conception du corpus décrite dans la section 3.1. L'équilibre entre signaux physiques et émotionnels dans ce corpus ne reflète pas la distribution naturelle observée en pratique : une expérience à venir sur le corpus Tamalou permettra de tester le pipeline sur une répartition plus réaliste et naturelle. Un autre biais est lié aux modèles de LLM eux-mêmes. Nous avons travaillé avec Mistral, un LLM généraliste non spécialisé dans les thèmes abordés. Malgré cela, Mistral a réussi à extraire (totalement ou partiellement) les symptômes et maladies dans la plupart des cas, même sans ajustement spécifique au domaine. D'autres modèles (Nemotron, Claude) ont présenté des performances inférieures pour extraire les informations pertinentes, ce qui peut s'expliquer par l'utilisation du même prompt que pour Mistral, sans adaptation spécifique à chaque famille de modèles.

5.2. Chaîne de valeur clinique

Les résultats présentés ici constituent une première étape vers une chaîne de valeur clinique plus large. Cette approche proposée ici constitue une première étape d'un dispositif plus large, qui intégrera progressivement l'analyse des sentiments, le traitement du langage naturel clinique et l'évaluation par des modèles d'IA avancés, afin de convertir les observations informelles en données directement utilisables et exploitables par les équipes soignantes. Une fois validés, les signaux faibles extraits par le pipeline seront structurés et priorisés selon les quatre dimensions du ZBI (charge personnelle, compétence perçue, relations interpersonnelles et attentes sociales), offrant ainsi une taxonomie cliniquement interprétable, fondée sur un instrument validé. Aucune de ces étapes ne serait possible sans les méthodes d'extraction développées dans cet article.

L'évaluation présentée vise à démontrer la faisabilité de l'extraction de signaux faibles à partir de messages, en utilisant des méthodes prêtes à l'emploi, facilement intégrables dans des applications collaboratives comme Tamalou. À long terme, notre objectif principal est d'utiliser les résultats structurés de cette analyse pour générer des notes compatibles avec le standard FHIR, destinées à être intégrées au Dossier Médical Partagé (DMP). Cela permettra de traduire les observations informelles des aidants en informations cliniques exploitables par les équipes soignantes.

Concrètement, nous pouvons envisager des fonctionnalités telles que des tableaux de bord, des seuils d'alerte configurables et des résumés synthétiques pour les équipes médicales. Leur conception, au même titre que la validation de la pertinence des données extraites, nécessitera une collaboration interdisciplinaire avec des médecins

généralistes, des gériatres et des experts en gestion de la douleur. Cette intégration répond directement au manque identifié par Greenhalgh et al. (2009) concernant l'incapacité des dossiers médicaux structurés à capturer les signaux du quotidien.

5.3. Considérations éthiques et gouvernance des données

Cette recherche implique l'accès à des données sensibles concernant des personnes vulnérables et leurs familles. Le corpus actuel a été constitué à partir de données publiques issues de forums de santé, ce qui, selon la réglementation française (RGPD, Art. 89), ne nécessite pas d'autorisation éthique institutionnelle, dans la mesure où : (i) tous les verbatims ont été intégralement anonymisés avant traitement, avec suppression systématique des informations permettant une identification ; (ii) les données n'ont pas été collectées sur la plateforme Tamalou, mais sur des forums publics ouverts ; (iii) aucune réidentification individuelle n'est possible à partir des résultats traités. La prochaine phase de la recherche, impliquant des messages d'utilisateurs réels de Tamalou, sera menée sous un protocole éthique formellement approuvé : La participation sera strictement volontaire (opt-in) ; les familles bénéficieront d'une transparence totale sur l'utilisation, le stockage et l'analyse des données ; et tous les traitements seront effectués en local (sur les infrastructures de l'Université de Reims), sans transfert des messages privés vers des serveurs tiers. Ce protocole sera soumis pour validation au comité d'éthique de l'Université Paris 1 Panthéon-Sorbonne avant toute collecte de données.

Nous sommes également conscients du risque de perception de surveillance par les familles : la méfiance envers l'analyse automatisée de messages privés est une préoccupation majeure. Nous y répondons par une communication claire sur l'utilisation des données et une conformité totale au RGPD et à l'EU AI Act. Enfin, pour éviter la surcharge informationnelle tant pour les aidants que pour les professionnels de santé, des critères de priorisation et des seuils d'alerte configurables seront mis en place.

6. Conclusions et Perspectives

Dans cet article, nous avons présenté une méthodologie visant à structurer et exploiter les signaux faibles, notamment la référence à des symptômes sur des conversations familiales informelles dans un contexte d'accompagnement. Notre approche s'inscrit dans le cadre de l'Action Design Research (ADR) (Sein et al., 2011) et s'appuie sur trois théories fondamentales : la charge des aidants (Tang et al., 2018), la dynamique des émotions (Kuppens et Verduyn, 2017), et la conscience de groupe dans le domaine du CSCW (Dourish et Belotti, 1992). Conformément à Gregor et Hevner (2013), notre contribution se situe à l'intersection de l'amélioration et de l'innovation. Nos premiers résultats sont prometteurs : l'extraction des symptômes atteint près de 70 % d'inférences correctes, celle des maladies 50,7 % de précision, tandis que le taux d'hallucinations reste faible (0,695 par ligne). Ces résultats démontrent la faisabilité de notre approche, tout en identifiant des pistes claires d'amélioration.

Les travaux futurs s'articuleront autour de quatre axes : (1) Tester le pipeline sur le corpus Tamalou, incluant des conversations complètes au sein d'un même cercle familial, avec des données longitudinales ; (2) Intégrer l'analyse des sentiments et des règles grammaticales (*e.g.* MoLeSy (Zhang et al., 2022) pour mieux traiter le langage informel ; (3) Explorer différentes familles de modèles pour les rôles d'inférence et de jugement, afin de réduire les biais de circularité ; (4) Associer des professionnels de santé (médecins généralistes, gériatres) à la co-conception de critères de priorisation des alertes et de formats de sortie compatibles avec le Dossier Médical Partagé (DMP).

À long terme, l'objectif principal de ce travail est de créer un pont entre les connaissances familiales et l'expertise médicale, avec des enjeux multiples : améliorer la qualité de vie des patients grâce à des ajustements précoces ; optimiser le suivi pour le système de santé ; reconnaître le rôle des familles dans l'accompagnement. Nous sommes convaincus que les familles constituent des remarquables capteurs humains pour des données cliniques, et que les nouvelles technologies, dont les grands modèles de langage (LLM), utilisés avec empathie, peuvent donner une voix à leurs observations silencieuses.

Remerciements. *Les travaux du premier et du deuxième auteur ont été financés, respectivement, par Tamalou et l'Université Paris 1 Panthéon-Sorbonne.*

Bibliographie

Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jindi, D., Naumann, T., E Mcdermott, M. (2019). « Publicly Available Clinical BERT Embeddings ». Proceedings of the 2nd Clinical Natural Language Processing Workshop, pp. 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics (2019). DOI : 10.48550/arXiv.1904.03323.

Branco, F., Gonçalves, R., Martins, J., Bessa, J. Et Baptista, A. (2016). « Computer Supported Co-operative Work—Exploratory Study on CSCW and Groupware Technologies and Its Applicability in the Health Area ». In Á. Rocha, A. M. Correia, H. Adeli, L. P. Reis, M. Mendonça Teixeira (ds.), New Advances in Information Systems and Technologies, Advances in Intelligent Systems and Computing, vol. 445. Springer International Publish-ing (2016). DOI : 10.1007/978-3-319-31307-8_40.

Brown, T., Mann, B., Ryder, N., et al. (2020). « Language models are few-shot learners ». 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada. Advances in neural information processing systems, vol. 33. https://proceedings.neurips.cc/paper_files/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

Diaz Ochoa, J.G., Layer, N., Mahr, J., Mustafa, F.E., Menzel, C.U., Müller, M.; Schilling, T., Illerhaus, G., Knott, M., Krohn, A. (2025). « Optimized BERT-Based NLP Outperforms Zero-Shot Methods for Automated Symptom Detection in Clinical Practice ». Prepublication, Health Informatics, April 22th 2025. DOI : 10.1101/2025.04.21.25326037

Dourish, P., Bellotti, V. (1992). « Awareness and Coordination in Shared Workspaces ». Proceedings of ACM Conference on Computer-Supported Cooperative Work (CSCW'92), ACM Press, pp. 107–114. DOI : 10.1145/143457.143468

Ellis, C.A., Simon J.G., Rein, G. (1991). « Groupware: Some Issues and Experiences ». Communications of the ACM, vol. 34, n° 1, pp. 39-58. DOI : 10.1145/99977.99987

Greenhalgh, T., T., Potts, H. W., Wong, G., Bark, P., Swinglehurst, D. (2009). « Tensions and Paradoxes in Electronic Patient Record Research ». Milbank Q., vol. 87, n° 4, pp. 729–788.

Gregor, S., Hevner, A.R. (2013). « Positioning and Presenting Design Science Research for Maximum Impact ». MIS Quarterly, vol. 37, n° 2, article n° 337355. DOI: 10.25300/MISQ/2013/37.2.01.

Grudin, J. (1994). « Computer-supported cooperative work: history and focus ». Computer, vol. 27, n° 5, pp.19–26.

Gu, J., Jiang, X., Shi, Z. et al. (2026). « A survey on LLM-as-a-Judge ». The Innovation, Jan. 2026, 101253. Elsevier. DOI : 10.1016/j.xinn.2025.101253.

Klasnja, P., Pratt, W. (2012). « Healthcare in the pocket ». Journal of Biomedical Informatics, vol. 45, n° 1, pp. 184–198.

Kuppens, P., Verduyn, P. (2017). « Emotion dynamics ». Current Opinion in Psychology, vol. 17, pp. 22–26.

Lee, S. A., Wu, A., Chiang, J.N. (2025). « Clinical ModernBERT: An efficient and long context encoder for biomedical text ». Prepublication, arXiv, 2025. https://arxiv.org/abs/2504.03964

Ji, Z., Lee, N., Frieske, R. et al., (2023) « Survey of Hallucination in Natural Language Generation », ACM Comput. Surv., vol. 55, nº 12, p. 1-38, déc. 2023, DOI : 10.1145/3571730.

Liang, P., Bommasani, R., Lee, T. et al., « Holistic Evaluation of Language Models », Transactions on Machine Learning Research, vol. 08/2023, https://openreview.net/pdf?id=iO4LZibEqW.

Liévin, V., Hother, C. E., Motzfeldt, A. G., Winther, O. (2024). « Can large language models reason about medical questions? », Patterns, vol. 5, nº 3, p. 100943, mars 2024, DOI : 10.1016/j.patter.2024.100943.

Lester, B., Al-Rfou, R., Constant, N. (2021). « The Power of Scale for Parameter-Efficient Prompt Tuning ». Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 3045-3059. DOI : 10.18653/v1/2021.emnlp-main.243

Mejia, D.A., Morán, A.L., Favela, J. (2007). « Supporting Informal Co-Located Collaboration in Hospital Work ». In: Haake, J.M., Ochoa, S.F., Cechich, A. (eds), Groupware: Design, Im-plementation, and Use, CRIWG 2007, Lecture Notes in Computer Science, vol. 4715, Springer Berlin Heidelberg, pp. 255-270. DOI : 10.1007/978-3-540-74812-0_20

Montgomery, R.J.V., Gonyea, J.G., Hooyman, N.R. (1985). « Caregiving and the Experience of Subjective and Objective Burden ». Family Relations, vol. 34, n° 1, pp. 19–26.

Nori, H., King, N., McKinney, S. M., Carignan, D., Horvitz, E. (2023). « Capabilities of GPT-4 on Medical Challenge Problems ». arXiv. https://doi.org/10.48550/ARXIV.2303.13375

Romanov, A. Shivade, C. (2018). « Lessons from Natural Language Inference in the Clinical Domain ». Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium. Association for Computational Linguistics, pp. 1586-96. DOI : 10.18653/v1/D18-1187.

Saha, S., Levy, O., Celikyilmaz, A., Bansal, M., Weston, J., & Li, X. (2024). « Branch-Solve-Merge Improves Large Language Model Evaluation and Generation ». Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol. 1: Long Papers), pp. 8352-8370. DOI : 10.18653/v1/2024.naacl-long.462

Sein, M.K., Henfridsson, O., Purao, S., Rossi, M., Lindgren, R. (2011). « Action design research ». MIS Quarterly, vol. 35, n° 1, pp. 37–56.

Si, Y. Wang, J., Xu, H., Roberts, K. (2019). « Enhancing Clinical Concept Extraction with Contex-tual Embeddings ». Journal of the American Medical Informatics Association, vol. 26, n° 11, pp. 1297-304. DOI : 10.1093/jamia/ocz096.

Singhal, K., Azizi, S., et al. (2023). « Large language models encode clinical knowledge ». Nature, vol. 620, pp. 172-180. DOI : 10.1038/s41586-023-06291-2

Smriti, D., Wang, L., Huh-Yoo, J. (2024). « Emotion Work in Caregiving: The Role of Technology to Support Informal Caregivers of Persons Living With Dementia ». Proceedings of the ACM on Human-Computer Interaction, vol. 8, issue CSCW1, article n° 48, pp. 1-34. DOI : 10.1145/3637325.

Sun, W. Rumshisky, A. Uzuner, O. (2013). « Evaluating Temporal Relations in Clinical Text: 2012 I2b2 Challenge ». Journal of the American Medical Informatics Association, vol. 20, n° 5, pp. 806-13. DOI : 10.1136/amiajnl-2013-001628

Tang, C., Chen, Y., Cheng, K., Ngo, V., Mattison, J.E. (2018). « Awareness and Handoffs in Home Care: Coordination among Informal Caregivers ». Behaviour & Information Technology, vol. 37, n° 1, pp. 66-86. DOI : 10.1080/0144929X.2017.1405073.

Wei, J., Wang, X., Schuurmans, D., et al. (2022). « Chain-of-thought prompting elicits reasoning in large language models ». Proceedings of the 36th Int. Conf. on Neural Information Processing Systems (NIPS '22), Article 1800, pp. 24824–24837.

Zarit, S.H., Reever, K.E., Bach-Peterson, J. (1980). « Relatives of the impaired elderly: Correlates of feelings of burden ». The Gerontologist, vol. 20, n° 6, pp. 649–655.

Zhang, B., Zhang, H., Shang J., Cai, J. (2022). « An Augmented Neural Network for Sentiment Analysis Using Grammar ». Frontiers in Neurorobotics, 16, 01 July 2022, paper n° 897402. DOI : 10.3389/fnbot.2022.897402

Zheng, L., Chiang, W.-L., Sheng, Y. et al. (2023). « Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena ». Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23). Article n°. 2020, pp. 46595 – 46623. https://huggingface.co/papers/2306.05685

An Evidence Model for Trustworthy Forecast Delivery in Multi-Site PV Systems

Leila Sakli[1] , Seifeddine Ben Elghali[1]

1. Aix Marseille Univ, CNRS, LIS, Marseille, France;
{leila.sakli,seifeddine.benelghali}@lis-lab.fr

RESUME. *La prévision photovoltaïque (PV) multi-sites est de plus en plus délivrée comme un service opérationnel intégré dans les systèmes d'information. La précision prédictive seule est insuffisante : les parties prenantes doivent pouvoir retracer quels modèles ont été utilisés, comment une prévision a été sélectionnée, et comment les conditions dégradées ont été gérées. Nous proposons un modèle d'évidence structurant les faits d'exécution en trois catégories : évidence d'exécution (QoS au niveau job), évidence de décision (sélection de modèles par site), et évidence d'échec (diagnostics typés et actions de continuité). Nous instancions le modèle dans un service réel et évaluons sa capacité à rendre la livraison de prévisions auditable et compréhensible sans accès aux paramètres internes des modèles.*

ABSTRACT. *Multi-site photovoltaic (PV) forecasting is increasingly delivered as an operational service embedded in information systems. Predictive accuracy alone is insufficient: downstream stakeholders must be able to trace which models were used, how a forecast was selected, and how degraded conditions were handled. We propose an evidence model structuring runtime facts into three categories: (i)* execution evidence *(job-level QoS), (ii)* decision evidence *(per-site model selection), and (iii)* failure evidence *(typed diagnostics and continuity actions). We instantiate the model in a multi-site PV delivery service and demonstrate lightweight evidence-quality indicators (completeness, diagnosability, continuity, timeliness, consistency), enabling auditable and interoperable traceability without requiring access to internal model parameters.*

MOTS-CLÉS : *Livraison de prévision, Modèle d'évidence, Confiance, Explicabilité au niveau processus, Prévision photovoltaïque.*

KEYWORDS: *Forecast delivery, Evidence model, Trust, Understandability, Process-level explainability, Photovoltaic forecasting.*

1. Introduction

Forecasting services are increasingly embedded into information systems that support operational decision-making (e.g., energy management and scheduling) (Pierro *et al.*, 2023). In such settings, the value of a forecast is not only determined by predictive accuracy but also by the ability of downstream stakeholders and systems to understand and trust how the forecast was produced and delivered (Kaur *et al.*, 2022). In operational forecasting services, delivery performance can be characterized using Quality of Service (QoS) metrics. Such metrics, widely studied in workflow and service systems (Cardoso et al., 2004), include latency, throughput, and reliability. Beyond accuracy, operational accountability requires distinguishing nominal operation from degraded conditions and supporting post-hoc justification of forecast-driven decisions (Rawal *et al.*, 2022). In this paper, we refer to *evidence* as runtime facts emitted by a delivery service to enable auditability and understandability of its behavior.

In practice, multi-site PV forecast delivery pipelines combine heterogeneous predictors (e.g., physics-based baselines and data-driven models) and execute under time constraints. Real-world operation exposes recurring issues often underrepresented in accuracy-focused studies, including common data issues (e.g., missing values, misalignment, or insufficient data) (Sakli and Ben Elghali, 2025). When such conditions occur, a pipeline may silently degrade (e.g., skip a model, switch to a fallback, or publish an uncalibrated output), which reduces transparency and can undermine trust even when numerical performance remains acceptable (Rawal *et al.*, 2022).

Existing work on explainability has largely emphasized model-level interpretability (e.g., feature attribution) (Rodrigues *et al.*, 2024). While valuable, it does not explain process-level behavior of a forecast delivery service : which models were considered, how they were compared, which one was selected for a given site, and what continuity actions were taken under failure. This gap is critical for multi-site operations where selection may vary per site and degraded operation must be explicit and auditable.

This work makes three contributions. First, we propose an evidence model for forecast delivery that structures runtime information into complementary components covering execution (QoS), decision-making, and failure handling. Second, we instantiate the model in a multi-site PV forecast delivery service that emits machine-readable artifacts (job reports, per-site comparison reports, and linked plots/exports) suitable for downstream information systems. Third, we provide an evidence-driven evaluation using operational artifacts, reporting job-level QoS, per-site model-selection outcomes, a failure taxonomy grounded in observed degradations, and qualitative examples illustrating divergence and fallback behavior.

The remainder of this paper is organized as follows. Section 2 positions the work through background and related work. Section 3 presents the Design Science research method. Section 4 formalizes the problem and requirements. Section 5 introduces the evidence model. Section 6 describes the system instantiation. Section 7 details the evaluation setup, and Section 8 reports results. Section 9 discusses implications and threats to validity. Finally, Section 10 concludes and outlines future work.

2. Background and Related Work

Our work lies at the intersection of operational forecasting services, explainability, and information-systems concerns such as auditability and interoperability. We position our contribution as *process-level explainability* for forecast delivery, rather than model-level interpretability. More broadly, trustworthy AI spans multiple dimensions (including robustness, privacy, fairness, and explainability) that motivate stronger accountability for AI-enabled services (Kaur *et al.*, 2022; Rawal *et al.*, 2022).

Multi-site PV forecasting has been extensively studied from the perspective of predictive performance, including heterogeneous predictors and, in some settings, ensemble strategies (Pierro *et al.*, 2016). In operational settings, however, forecasting is delivered as a service under time constraints and recurring data issues, which makes delivery-time traceability and degraded-mode transparency critical for downstream stakeholders and integrating systems.

Multi-site PV forecasting faces recurring operational challenges including missing or invalid inputs, temporal misalignment, and insufficient historical data. Ensemble methods that combine multiple predictors can improve accuracy (Pierro *et al.*, 2016), but their operational behavior (e.g., which models were evaluated, how they were ranked, and which one was selected) is often opaque. Context-aware stacking approaches have been proposed to handle heterogeneous site conditions (Sakli and Ben Elghali, 2025), but do not explicitly emit evidence of model-selection decisions.

A substantial literature addresses explainability at the model level (e.g., feature attribution), enabling insight into why a particular model produced a prediction ; a prominent example is SHAP (Lundberg and Lee, 2017). Explainability techniques for time series forecasting have also focused on feature importance and model interpretation (Rodrigues *et al.*, 2024). While valuable, such techniques do not explain how a delivery pipeline behaved at runtime in multi-model settings : which candidates were evaluated, how they were compared, which model was selected per site, and what continuity actions were taken when inputs or models failed. Runtime explainability frameworks for end-to-end ensemble ML serving move toward execution-time transparency (Nguyen *et al.*, 2024), but have not yet been widely explored in the context of multi-site forecasting with explicit degraded-mode semantics. ML monitoring work highlights the importance of explainability in production settings. Modern observability and MLOps practices (logs, metrics, traces) are essential to operate ML pipelines, but they do not provide, by themselves, a stable and semantically explicit account of forecast delivery decisions that can be consumed by downstream information systems. First, observability signals are primarily implementation-centric : they are high-volume, pipeline-specific, and often evolve with instrumentation choices, which makes long-term auditing and cross-system interoperability brittle. Second, they are typically component-health oriented rather than decision oriented : they may indicate that a training step failed, but they do not directly reconstruct, at entity/site scope, which candidate models were considered, how they were ranked, and which forecast was ultimately published. Third, observability rarely encodes degraded-mode semantics :

when inputs are missing or a predictor fails, stakeholders need an explicit statement of what continuity action was taken (e.g., best-available selection vs. fallback) and why, to avoid silent degradations. Workflow QoS metrics provide useful execution-level signals (e.g., runtime and reliability) (Cardoso *et al.*, 2004), but do not capture forecasting-specific decision traceability or failure/continuity semantics.

We therefore introduce a minimal *evidence contract* that complements observability with decision-level semantics : job-level execution QoS ; per-entity candidate set, metrics, ranking, and delivered `best_model` ; and typed failure diagnostics paired with an explicit continuity action. Because evidence uses stable identifiers (`job_id`, `entity_id`) and machine-checkable consistency constraints, it supports automated ingestion, post-hoc accountability, and evidence-quality assessment without requiring access to internal model parameters. In short, observability supports debugging ; evidence supports interoperable auditing and trustworthy delivery.

3. Research Method (Design Science)

We follow a Design Science Research (DSR) approach (Hevner and Chatterjee, 2010), which is appropriate when the research contribution is an artifact that addresses a practical problem and is evaluated through its utility and measurable properties in context. In our case, the central artifact is an evidence model for forecast delivery, together with an evidence-driven evaluation protocol.

The proposed artifact consists of two complementary components. **Artifact A** defines an evidence model, i.e., a minimal, structured representation of runtime facts emitted by the forecast delivery service, covering execution QoS, decision traceability (candidate models, ranking, selected `best_model`, and metrics), and failure transparency with declared continuity actions. **Artifact B** provides an evidence-driven evaluation protocol, i.e., a lightweight evaluation methodology that measures trustworthy delivery using the emitted artifacts (job reports, per-site comparison reports, and forecast exports), without requiring additional simulations beyond normal operational runs. The artifact is designed to be model-agnostic and compatible with heterogeneous predictors, enabling adoption in existing multi-model forecasting pipelines.

The design objectives are derived from the requirements in Section 4. Specifically, the artifact must (i) provide measurable QoS signals (R1), (ii) allow reconstruction of model-selection decisions (R2), (iii) make degraded operation explicit and diagnosable while maintaining continuity (R3), and (iv) remain lightweight and interoperable for downstream information systems (R4).

Following DSR practice, we evaluate the artifact by demonstrating that it enables trustworthy and understandable service behavior through measurable evidence properties and empirical observations. The evaluation uses operational artifacts produced by the system along four dimensions : QoS feasibility (runtime, throughput, input retrieval success/failure), decision traceability (standardized metrics, model rankings, and `best_model` decisions with linked plots), failure transparency and continuity (typed

error records paired with an explicit action), and qualitative interpretability (forecast exports illustrating divergence and fallback behavior). In addition to conventional predictive metrics (RMSE, MAE, MAPE, correlation), we assess the quality of the emitted evidence using lightweight indicators defined later in the paper.

Relevance is ensured by grounding the artifact in operational constraints and failure modes observed in a multi-site deployment setting. Rigor is addressed by specifying a minimal evidence schema with explicit field semantics, evaluating the artifact through reproducible artifacts and metrics extracted from system outputs, and clearly separating the proposed evidence model (research contribution) from the underlying forecasting predictors (implementation context).

4. Problem and Requirements

We study multi-site PV forecast delivery as an information system problem rather than a pure prediction task. In operational settings, forecasting pipelines typically combine heterogeneous predictors (e.g., physics-based and data-driven models) and are executed under time constraints to deliver a day-ahead product. However, real-world conditions frequently include degraded inputs (missing values, delayed data, time-index misalignment) and partial unavailability of models. In such cases, accuracy alone is insufficient to establish trust : downstream stakeholders and integrating systems need to understand what the service actually executed, which model was selected, why some candidates failed, and what continuity action was taken.

As argued in the introduction, model-level explainability does not address process-level transparency of a delivery service. We derive four requirements for a trustworthy and understandable forecast delivery service.

R1 — Execution QoS (measurable service behavior). The service shall expose execution facts that allow stakeholders to reason about operational feasibility and reliability, including runtime duration, throughput (number of processed sites), and success/failure of data retrieval for key inputs (PV measurements, weather features). These QoS indicators are required to assess whether forecast delivery is compatible with operational deadlines.

R2 — Decision traceability (explainable selection). The service shall represent model-selection decisions in a reproducible manner. For each site, it must be possible to identify the set of evaluated candidate predictors, the ranking criterion, the selected `best_model`, and the associated performance metrics used to support this decision. Decision traceability is required for explainability at the process level and for integration into downstream decision workflows.

R3 — Failure transparency and continuity (degraded operation is explicit). When a candidate predictor cannot be trained or evaluated (e.g., missing/invalid inputs, time-index mismatch, insufficient training data), the service shall emit explicit failure

evidence (typed diagnostic and message) and declare the continuity action taken (skip affected model, select best among remaining, or fallback). This requirement prevents silent degradations and supports post-hoc diagnosis.

R4 — Low overhead and interoperability (IS-ready artifacts). Evidence emission shall be lightweight and interoperable. Artifacts should be machine-readable (e.g., JSON) with stable field semantics, enabling automated ingestion by downstream information systems, while remaining concise enough to avoid penalizing the delivery path.

Together, R1–R4 define a design target where forecast delivery is not only numerically accurate but also operationally auditable and understandable.

5. Evidence Model

This section introduces the proposed evidence model : a minimal, machine-readable representation of runtime facts that supports trustworthy forecast delivery in multi-site PV systems. The model is designed to expose (i) execution Quality of Service (QoS), (ii) model-selection decisions, and (iii) explicit failure diagnostics together with continuity actions. The evidence model is *model-agnostic* : it does not assume access to internal parameters of the predictors and can be attached to heterogeneous forecasting pipelines.

5.1. *Overview : three complementary evidence views*

We structure evidence into three complementary views that jointly explain the observable behavior of a delivery service : (i) **Execution evidence** describes *what ran* and *how it ran* at job scope (timing, throughput, and input retrieval success/failure), (ii) **Decision evidence** describes *what was decided* at entity/site scope (evaluated candidates, ranking, selected model, and metrics), and (iii) **Failure evidence** describes *what went wrong* and *what was done* (failure type/message, impacted model(s), and declared continuity action). This separation supports both automated monitoring (via stable structured fields) and human inspection (via concise summaries and linked artifacts), as shown in Figure 1.

5.2. *Minimal schema : job-level and entity-level evidence*

Table 1 summarizes the minimal fields and their semantics at job and entity scope. Fields are grouped by scope : job-level evidence captures workflow-wide facts for a multi-site run, while entity-level evidence captures per-site decisions and outcomes. The schema can be serialized as JSON and extended with optional metadata when available, while keeping core semantics stable for downstream ingestion.

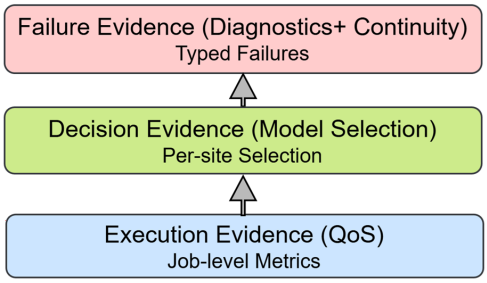


FIGURE 1. *Layered architecture of the evidence model.*

5.3. *Field semantics and interoperability conventions*

The evidence record is designed to be IS-ready and interoperable. Job-level fields provide monitoring signals at workflow scope (e.g., whether a job met operational expectations in terms of duration and input retrieval), while entity-level fields enable post-hoc reconstruction of model selection and degraded operation for each site. Interoperability is achieved through : (i) stable identifiers (`job_id`, `entity_id`) linking job and per-site records ; (ii) machine-readable serialization (JSON for evidence records, CSV/JSON for forecast exports) ; and (iii) optional pointers to human-facing artifacts (plots, reports) referenced from `artifacts`. This design supports auditing without requiring access to proprietary model parameters.

5.4. *Selection semantics and publication validity*

Decision evidence captures how the service determines which forecast is ultimately published for each entity. For a given entity e, let M denote the set of available candidate predictors and let $M^{+}(e) \subseteq M$ be the subset that successfully produced evaluation results. The service records, for each $m \in M^{+}(e)$, standardized metrics, together with the resulting ranking and the selected `best_model`. The selection policy is configurable ; in all cases, evidence makes the evaluated candidates, criterion/metrics, ranking, and final decision auditable.

In our instantiation, the Ensemble predictor follows a constrained ensemble-learning scheme (Sakli and Ben Elghali, 2025), combining base predictions through a ridge-regularized linear meta-learner (constrained Ridge regression) complemented by a context-aware stacking mechanism, and is treated as one candidate among the evaluated predictors. The delivered `best_model` follows the configured selection policy based on the available per-model metrics and operational constraints. Independently of the selection outcome, the published forecast is passed through post-prediction validation ;

TABLEAU 1. *Minimal evidence schema for forecast delivery.*

Scope	Field	Semantics
Job	`job_id`	Identifier for the run/workflow instance (e.g., daily update or evaluation batch).
Job	`workflow`	Workflow type (e.g., `daily_update`, `evaluation_batch`).
Job	`start_time`, `end_time`	Execution time boundaries used to compute duration and trace ordering.
Job	`duration_s`	Total runtime in seconds for the workflow execution.
Job	`processed_entities`	Number of sites/entities processed in the job.
Job	`input_status`	Success/failure counts for PV and weather retrieval (and optional error summaries).
Entity	`entity_id`	Unique site identifier for which decisions and forecasts are produced.
Entity	`horizon_start`, `horizon_end`	Forecast horizon boundaries (timestamps) for the delivered product.
Entity	`candidates`	Candidate predictors considered for the entity (e.g., Physical, XGBoost, LSTM, Ensemble).
Entity	`metrics`	Per-model metrics (RMSE, MAE, MAPE, correlation, and sample count `n`) when available.
Entity	`ranked_models`	Ordered list of successful candidates from best to worst according to the selection rule.
Entity	`best_model`	identifier of the model ultimately delivered for publication (policy outcome), possibly reflecting best-available or fallback under degraded conditions.
Entity	`failure`	Optional failure record with `failure_type`, `message`, and `impacted_models`.
Entity	`action`	Continuity action in {`nominal`, `best-available`, `fallback`}.
Entity	`artifacts`	Pointers to artifacts (e.g., report path, plot path, forecast export path).

if validation rejects an output, the service records the corresponding failure/continuity evidence and triggers a best-available or physics-based fallback action.

5.5. *Failure semantics and continuity actions*

Failures are represented as first-class structured objects rather than unstructured log lines. A failure record minimally includes a coarse `failure_type` (e.g., `nan_input`, `timestamp_mismatch`, `no_data`), a human-readable `message`, and the `impacted_models`. Crucially, failures are paired with an explicit continuity action describing how delivery continuity was maintained : (i) `nominal` (no failure affecting selection/publication), (ii) `best-available` (some candidates failed but a best model is selected among remaining), or (iii) `fallback` (no trained model available or insuffi-

cient data ; a fallback output is published). This pairing prevents silent degradations and enables downstream systems to distinguish nominal from degraded operation.

A compact evidence-driven procedure for each entity e and horizon H is :

1) Record job-level execution evidence (timestamps, duration, input retrieval success/failure).

2) For each candidate $m \in M$: if training/evaluation fails, record `success=false` with an error message ; else record metrics.

3) Compute $M^+(e) = \{m \in M : \text{success}(m) = \text{true}\}$ and select `best_model` according to the configured selection policy (based on the available per-model metrics), and record the resulting ranking and decision evidence.

4) If $M^+(e) = \emptyset$ (or insufficient data), publish fallback and set `action=fallback` ; otherwise publish `best_model` and set `action=nominal` or `best-available`.

5) Store decision evidence (`ranked_models`, `best_model`, metrics) and artifact pointers (report/plot/export).

5.6. *Evidence-quality indicators*

Beyond forecast accuracy, we assess the quality of the emitted evidence. Let S be the set of processed entities (sites) in a run and let $E(e)$ denote the emitted evidence record for entity $e \in S$. Let Req denote the set of required fields for a workflow.

Completeness/Coverage. Evidence completeness is the fraction of entities for which all required fields are present :

$$Comp = \frac{1}{|S|} \sum_{e \in S} \mathbf{1}\big[Req \subseteq \mathit{fields}(E(e))\big]. \quad [1]$$

Diagnosability/Granularity. Let $F \subseteq S$ be the subset of entities where at least one candidate model fails (a non-empty failure record is applicable). Diagnosability measures whether failures are sufficiently described for root-cause analysis :

$$Diag = \frac{1}{|F|} \sum_{e \in F} \mathbf{1}\big[\texttt{failure\_type} \wedge \texttt{message} \wedge \texttt{impacted\_models} \text{ present in } E(e)\big]. \quad [2]$$

Continuity. Let $D \subseteq S$ be the subset of degraded entities (some model failure or insufficient data). Continuity measures the ability to still produce a declared outcome :

$$Cont = \frac{1}{|D|} \sum_{e \in D} \mathbf{1}\big[\texttt{best\_model} \text{ defined} \vee \texttt{action=fallback}\big]. \quad [3]$$

Timeliness. Timeliness assesses whether evidence is produced early enough to support operational use. Let $t_{emit}(e)$ be the timestamp at which the entity evidence is finali-

zed, and let $t_{pub}(e)$ denote the forecast publication time (or the job `end_time` when publication is synchronous). A minimal indicator is :

$$Time = \frac{1}{|S|} \sum_{e \in S} \mathbf{1}\big[t_{emit}(e) \leq t_{pub}(e)\big], \quad [4]$$

which can be refined when an explicit delivery deadline is available (e.g., $t_{emit}(e) \leq t_{deadline}$).

Consistency. Consistency checks alignment between declared actions and observable fields. A minimal set of verifiable constraints is : (i) if `action=nominal` then `failure` is empty and `best_model` is defined ; (ii) if `action=best-available` then `best_model` is defined and at least one candidate failure is recorded ; (iii) if `action=fallback` then $M^{+}(e) = \emptyset$ (or `no_data`) and a fallback warning/flag is present in publication metadata (when available). Let $C \subseteq S$ be the subset of entities for which `action` is present ; then :

$$Cons = \frac{1}{|C|} \sum_{e \in C} \mathbf{1}\big[\text{consistency constraints hold for } E(e)\big]. \quad [5]$$

Finally, we report job-level QoS derived from evidence. For a job of duration T seconds processing $|S|$ entities, throughput is :

$$Throughput = \frac{|S|}{T} \quad \text{(entities/s)}, \quad [6]$$

and when input retrieval success/failure counts are available, availability is :

$$Avail = \frac{succ}{succ + fail}. \quad [7]$$

6. System Instantiation

We instantiate the evidence model in an operational multi-site PV forecast delivery service. The goal of this instantiation is not to introduce new predictors, but to demonstrate that evidence can be emitted as part of a real delivery pipeline and consumed by downstream information systems for monitoring, auditing, and accountability.

The service operates on a set of PV sites (entities) and produces day-ahead forecasts at hourly resolution. In our instantiation, base predictors include a physics-based model, an XGBoost model, and a two-layer LSTM. In addition, an Ensemble predictor follows a constrained ensemble-learning scheme (Sakli and Ben Elghali, 2025) : base predictions are fused using a ridge-regularized linear meta-learner (constrained Ridge regression) and complemented by a context-aware stacking mechanism. Delivered forecasts are then passed through a post-prediction filtering/validation step that enforces output validity and continuity ; if filtering cannot yield a valid forecast, the service

triggers a best-available or physics-based fallback and records the corresponding failure/continuity evidence.

Two operational workflows are supported. A *daily update* (delivery) job runs under operational time constraints to retrieve the latest PV measurements and meteorological inputs and publish forecasts for multiple entities together with minimal publication metadata (`entity_id`, horizon timestamps, model identifier, and warning/fallback flags when applicable). An *offline evaluation* job is compute-intensive and executed less frequently ; it compares candidate predictors per entity, computes standardized performance metrics, produces a ranked list and a `best_model` decision according to the configured policy, and generates per-site comparison artifacts (reports and plots). These artifacts can be consumed by the delivery workflow and also serve as evidence for post-hoc auditing and diagnosis.

To ensure operational continuity, the publication layer enforces output validity and applies two policies. If a selected output is invalidated (or a predictor fails) but at least one valid candidate output remains, the service publishes the best available forecast and declares `action=best-available`. If no valid forecast can be produced for an entity (e.g., due to insufficient data or invalid outputs), the service falls back to the physics-based baseline and declares `action=fallback`, attaching a warning flag indicating that the output may be uncalibrated. All evidence artifacts are stored in interoperable formats (JSON for reports ; CSV/JSON for forecast exports) and linked through stable identifiers (`job_id`, `entity_id`) to support traceability across job-level execution, site-level decisions, and degraded-operation handling.

7. Evaluation Setup

This section describes the empirical material and protocol used to evaluate the proposed evidence model and its usefulness for trustworthy forecast delivery. The evaluation is intentionally lightweight : rather than running additional simulations, we rely on operational artifacts that are already produced by the system during normal delivery and evaluation runs.

We consider a multi-site PV forecast delivery service operating on a day-ahead horizon at hourly resolution. The system runs (i) a daily update job, which retrieves inputs and delivers forecasts under operational time constraints, and (ii) an offline evaluation job, which compares candidate predictors and produces per-site model-selection artifacts. Both workflows emit machine-readable evidence records and linked artifacts (e.g., JSON reports and comparison plots), which can be inspected by downstream stakeholders and integrating information systems.

Evaluation relies on the following evidence sources. A *daily update report* provides job-level execution QoS evidence (processed sites, success/failure counts for PV and weather retrieval, and total runtime). An *evaluation job report* enumerates evaluated entities and, for each entity, the ranked candidate models, the selected `best_model`, and any training/evaluation failures with explicit error messages. *Per-site comparison*

reports provide standardized performance metrics per candidate predictor, the resulting model ranking, and pointers to associated comparison plots. *Forecast exports* (CSV/JSON) provide hourly series for selected dates, enabling qualitative visualization of inter-model divergence and fallback behavior. When available, training reports document training behavior and cost without requiring re-execution.

For each site, the system evaluates heterogeneous predictors (physics-based, XGBoost, and a two-layer LSTM) and, when enabled, an Ensemble predictor that fuses base predictions through constrained Ridge regression and a context-aware stacking mechanism. Delivered outputs are validated through post-prediction filtering, which may trigger best-available selection or fallback to the physics-based baseline under invalid outputs.

We report two complementary classes of metrics. *Forecast accuracy* is reported per model and site using RMSE, MAE, MAPE, Pearson correlation, and the evaluated sample count n, as provided by the per-site comparison reports. *Operational QoS* is reported at job level using the number of processed sites, input retrieval success/failure counts, and total execution time, as emitted by the daily update report and evaluation job report. To assess trustworthy delivery beyond accuracy, we compute evidence-quality indicators from emitted artifacts (completeness/coverage, diagnosability/granularity, continuity, timeliness, and consistency) as defined in Section 5. Finally, we use the failure evidence to derive a minimal failure taxonomy grounded in observed degradations and the declared continuity actions taken by the service.

8. Results

8.1. *Execution QoS and operational feasibility*

Table 2 contrasts two representative workloads from a sample run : a just-in-time daily delivery/update run and an offline evaluation batch run. Their execution times differ due to distinct purposes (time-constrained delivery vs. compute-intensive evaluation). In the daily update run (2025-06-30), the service processed 19 sites in 4.06 s with PV retrieval success 19/0. Meteo retrieval is monitored at job scope as provider availability ; 2/0 denotes two provider fetch attempts (primary and backup), both successful. This corresponds to an observed throughput of approximately 4.68 sites/s (0.214 s/site on average), supporting time-constrained delivery. The evaluation batch run (12129.65 s $\approx$ 3h22) reflects offline model evaluation and should not be interpreted as delivery-time performance. These job-level QoS signals provide auditable evidence of execution behavior and workflow feasibility (R1).

TABLEAU 2. *Job-level QoS evidence from operational reports.*

Job	Date/Scope	Sites	PV (s/f)	Meteo (s/f)	Duration
Daily update	2025-06-30	19	19 / 0	2 / 0	4.06 s
Eval. batch	multi-site	–	–	–	12129.65 s ($\approx$ 3h22)

8.2. *Evidence-backed model selection across sites*

Table 3 reports per-site model comparison outcomes for four representative sites selected from the 19-site run and extracted from `comparison_report` artifacts (RMSE, MAE, MAPE, correlation, and sample count), together with the delivered model identifier (`best_model`). Results show site-dependent selection : the Physical predictor is selected for three sites, while an LSTM predictor is selected for one. The artifacts also include an explicit ensemble status (OK/FAILED), making partial degradation visible. All metrics are computed on site-normalized PV power (per-unit scaling), enabling cross-site comparison. Overall, decision evidence supports post-hoc reconstruction of evaluated candidates, comparison criteria, and delivered models (R2). In Table 3, `best_model` denotes the delivered model under the configured selection policy, possibly overridden by post-prediction validation (best-available or fallback).

TABLEAU 3. *Per-site evidence-backed delivery outcome and performance metrics extracted from* `comparison_report` *artifacts.*

Entity (site)	Delivered model	n	RMSE	MAE	MAPE(%)	Corr.	Ensemble
TR-B1	Physical	433	0.01284	0.00559	6.70	0.9958	OK
TR-B9	Physical	433	0.02230	0.00977	14.16	0.9859	OK
TR-B12	Physical	433	0.01285	0.00623	7.64	0.9960	OK
SP-Farm	LSTM	414	0.50040	0.29940	44.79	0.9691	FAILED

8.3. *Failure taxonomy and transparency of degraded conditions*

Operational artifacts capture failure modes affecting training and/or inference. Table 4 summarizes a minimal taxonomy grounded in observed failures, together with supporting evidence fields and declared continuity actions. Three representative classes are illustrated : missing/invalid values (NaN), temporal misalignment/index mismatch, and absence of training data. For each class, degraded operation is made explicit through typed diagnostics and declared continuity actions (skip affected model and select among remaining, or fallback). This pairing prevents silent degradation and supports auditable continuity under faults (R3).

8.4. *Evidence-quality and overhead*

Beyond accuracy metrics, emitted artifacts support quality assessment along completeness/coverage, diagnosability/granularity, continuity, timeliness, and consistency, as defined in Section 5. Over the 19-site daily update run : Completeness = 0.975, Diagnosability = 1.000, Continuity = 1.000, Timeliness = 0.956, and Consistency = 1.000. Artifact emission is integrated in normal runs : the daily delivery job completes in 4.06 s while producing job-level and site-level artifacts, suggesting low overhead

TABLEAU 4. *Observed failure modes and corresponding transparency/continuity evidence.*

Failure type	Typical symptom	Evidence fields (examples)	Continuity action
Missing/invalid values (NaN)	Model training cannot proceed	`success=false`, `error="Input contains NaN"`	Skip affected model ; select best among remaining models
Temporal misalignment/index mismatch	Ensemble training fails due to timestamp mismatch	`success=false`, `error="... not in index"`	Skip ensemble ; select best among remaining models
No training data available	Training not performed ; fallback is used	`data_points=0`, `training_time=0`	Fallback strategy activated (baseline output)

while preserving time-constrained delivery. These artifacts are machine-readable (e.g., JSON/CSV) and linked via stable identifiers, enabling automated ingestion by downstream information systems without requiring access to internal model parameters (R4).

8.5. *Illustrative cases*

Two representative cases illustrate how evidence supports auditing. In a *nominal* case (e.g., Physical selected and ensemble OK), decision evidence provides candidates, metrics, ranking, and the selected `best_model` with linked artifacts. In a *degraded* case (e.g., ensemble FAILED), failure evidence records diagnostics and decision evidence documents selection among remaining candidates. In both cases, stakeholders can identify the delivered model, available alternatives, and whether operation was nominal or degraded with explicit continuity action.

9. Discussion and Threats to Validity

The proposed evidence model addresses a recurrent gap in operational forecasting : stakeholders often need to audit delivery behavior rather than interpret a specific predictor. The contribution is therefore not another forecasting method, but a minimal, semantically explicit representation of runtime facts that makes forecast delivery decisions reconstructible. Compared to ad hoc logging, the model provides stable field semantics, explicit links between execution QoS, per-site selection outcomes, and failure/continuity actions, and evidence-quality indicators that can be computed from artifacts. This supports downstream information systems by enabling automated ingestion, monitoring, and post-hoc accountability without requiring access to internal model parameters.

From an information-systems perspective, the separation into execution, decision, and failure evidence is critical. Execution evidence supports operational feasibility

analysis (e.g., whether a daily delivery run meets time constraints and input availability). Decision evidence enables per-entity traceability in heterogeneous multi-model settings, where a single global model choice is often insufficient. Failure evidence paired with declared continuity actions prevents silent degradation and provides diagnosable signals when a pipeline skips models, disables ensembles, or falls back due to missing data or misalignment. Together, these mechanisms make nominal versus degraded operation explicit and auditable.

Our instantiation and evaluation have several limitations. First, the empirical observations and failure taxonomy are grounded in a specific operational setting and a finite set of sites, inputs, and candidate predictors ; additional deployments may reveal further failure modes or different operational constraints. Second, the evidence-quality indicators proposed here are lightweight by design and emphasize completeness, diagnosability, continuity, timeliness, and consistency ; they do not replace comprehensive governance processes or formal verification. Third, the evaluation reports accuracy metrics and operational QoS but does not include a controlled user study ; conclusions about perceived trust or decision impact are therefore indirect.

Threats to validity follow the same structure. *Internal validity* may be affected if evidence artifacts are incomplete or inconsistent due to implementation errors ; we mitigate this by defining required fields, storing stable identifiers linking artifacts, and introducing consistency constraints that can be automatically checked. *Construct validity* concerns whether the evidence-quality indicators capture what stakeholders mean by trustworthy delivery ; we mitigate this by grounding the indicators in operational needs (auditability, degraded-mode transparency, continuity) and by pairing them with concrete artifacts (reports, plots, exports). *External validity* is limited by the single-domain focus (PV forecasting) and by environment-specific failure patterns ; however, the model is predictor-agnostic and can be applied to other forecasting services that require batch processing, per-entity model selection, and continuity under degraded inputs. Finally, *conclusion validity* depends on the representativeness of the observed runs ; we therefore treat the reported evidence as operational demonstrations rather than universal performance claims.

Overall, the discussion highlights a key point : trustworthy forecasting in information systems requires transparency about delivery processes (what executed, what was selected, what failed, and what continuity action was taken), in addition to predictive performance. The evidence model and evaluation protocol provide a lightweight basis for such process-level explainability and auditability in operational multi-site forecast delivery.

10. Conclusion and Future Work

This paper addressed trustworthy multi-site PV forecast delivery as an information-systems problem rather than only an accuracy problem. We proposed a minimal evidence model that structures runtime facts into complementary evidence components.

We instantiated the model in an operational delivery service that emits machine-readable artifacts and evaluated it using operational reports and per-site comparison artifacts.

Results show that the emitted evidence supports auditable monitoring of execution feasibility (e.g., daily delivery vs. offline evaluation workloads), reconstructible per-site selection decisions, and explicit transparency of degraded conditions through a minimal failure taxonomy linked to continuity actions.

Future work includes (i) mapping the evidence schema to provenance/observability standards, (ii) extending the failure taxonomy and automated checks, and (iii) conducting a user study with operational stakeholders to measure how evidence impacts trust, diagnosis time, and decision confidence.

Bibliographie

Cardoso J., Sheth A., Miller J., Arnold J., Kochut K., « Quality of service for workflows and web service processes », *Journal of Web Semantics*, vol. 1, nº 3, p. 281-308, 2004.

Hevner A., Chatterjee S., « Design science research in information systems », *Design Research in Information Systems*, Springer, 2010.

Kaur D., Uslu S., Rittichier K., Durresi A., « Trustworthy artificial intelligence : A review », *ACM Computing Surveys*, 2022.

Lundberg S. M., Lee S.-I., « A unified approach to interpreting model predictions », *Advances in Neural Information Processing Systems*, p. 4765-4774, 2017.

Nguyen M., Truong H.-L., Truong-Huu T., « Novel contract-based runtime explainability framework for end-to-end ensemble machine learning serving », *Proceedings of the IEEE/ACM International Conference on AI Engineering (CAIN)*, 2024.

Pierro M., Bucci F., Cornaro C., Maggioni E., Perotto A., Pravettoni M., Spada F., « Multi-model ensemble for day-ahead prediction of photovoltaic power generation », *Solar Energy*, vol. 134, p. 132-146, 2016.

Pierro M., Cornaro C., Perez R., « Forecasting services for renewable energy systems : State of the art and future directions », *Renewable and Sustainable Energy Reviews*, vol. 175, p. 113156, 2023.

Rawal A., McCoy J., Rawat D., Sadler B., Amant R., « Recent advances in trustworthy explainable artificial intelligence : Status, challenges, and perspectives », *IEEE Transactions on Artificial Intelligence*, 2022.

Rodrigues E., Baghoussi Y., Mendes-Moreira J., « Explainability feature selection framework application for LSTM multivariate time-series forecast self optimization », *Expert Systems*, 2024.

Sakli L., Ben Elghali S., « PV forecasting with constrained ensemble learning and context-aware stacking », *Proceedings of the 37th International Conference on Tools with Artificial Intelligence (ICTAI)*, IEEE, 2025.

Modélisaton conceptuelle des campagnes de désinformation

Sidbewendin Angélique Yameogo[1], Régis Fleurquin[1], Nicolas Belloir[1,2], Wassila Ouerdane[3]

1. *IRISA, Vannes, France*

2. *CReC St-Cyr, Académie Militaire de St Cyr Coëtquidan, Guer, France*

3. MISC, Centrale Supélec, Université Paris-Saclay, Gif sur Yvette, France

Contact : nicolas.belloir@irisa.fr

RÉFERENCE DE L'ARTICLE INTERNATIONAL. Cet article est une synthèse de l'article : Sidbewendin Angelique Yameogo, Régis Fleurquin, Nicolas Belloir, and Wassila Ouerdane. 2025. Conceptual Modeling of Disinformation Campaigns. In Conceptual Modeling: 44th International Conference, ER 2025, Poitiers, France, October 20–23, 2025, Proceedings. Springer-Verlag, Berlin, Heidelberg, 358–375. https://doi.org/10.1007/978-3-032-08623-5_19

La diffusion massive de contenus trompeurs sur les réseaux sociaux et les médias en ligne a profondément transformé les dynamiques de manipulation de l'information. Ces phénomènes ne se limitent plus à des fausses nouvelles isolées, mais prennent la forme de campagnes structurées, déployées sur plusieurs plateformes, mobilisant des acteurs humains et/ou des logiciels, et s'inscrivant dans des objectifs politiques ou géopolitiques à moyen et long terme. Face à cette évolution, les acteurs luttant contre la désinformation (analystes de renseignement, journalistes, …) peuvent s'aider de différents outils, allant de la détection automatique de fausses informations par apprentissage profond (Hu et al., 2022) à l'analyse de graphes sociaux (Lakzaei et al. 2024) pour identifier des comportements coordonnés. Récemment, des approches hybrides sont apparues (Chalehchaleh et al., 2024), combinant de l'analyse de contenu multi-formes (text, images, data) avec de l'analyse de dissémination. Cependant, ces solutions présentent des limites (Terp et Breuer, 2022)). D'une part, elles couvrent rarement l'ensemble des différentes dimensions d'une campagne de désinformation. D'autre part, elles reposent souvent sur des modèles opaques, difficiles à expliquer dans des contextes sensibles où la traçabilité et l'interprétation sont essentielles. Dans ce contexte, la modélisation conceptuelle apparaît comme une approche prometteuse pour structurer les connaissances, favoriser l'interopérabilité des outils et renforcer l'explicabilité des analyses.

La désinformation peut être définie comme la diffusion intentionnelle de contenus faux ou trompeurs dans le but d'induire en erreur (Starbird et al., 2019). Néanmoins, cette définition reste insuffisante pour rendre compte des campagnes de désinformation, caractérisées par leur organisation, leur coordination et leur inscription dans une stratégie globale d'influence. Ces campagnes impliquent des objectifs explicites, des acteurs multiples, des techniques de manipulation variées et des cibles précisément identifiées. Les approches existantes se répartissent principalement en deux catégories. Les méthodes centrées sur le contenu s'appuient sur le traitement automatique du langage naturel et l'apprentissage automatique pour détecter des indices linguistiques ou stylistiques de tromperie (Zhou et Zafarani, 2020). Bien qu'efficaces localement, elles considèrent souvent les messages de manière isolée. À l'inverse, les approches comportementales (Vargas et al., 2020 ou Phan et al., 2023) analysent les dynamiques de propagation et les interactions entre comptes, mais sans intégrer finement la sémantique des contenus. Les solutions hybrides tentent de combiner ces deux dimensions, mais souffrent d'un manque de cadre conceptuel commun, ce qui limite leur cohérence et leur interprétabilité.

Pour répondre à ces limitations, nous proposons un modèle conceptuel de données (MCD) servant de socle sémantique à l'analyse des campagnes de désinformation. Ce modèle a pour objectif de structurer et relier les informations issues de sources hétérogènes, tout en restant compréhensible pour des analystes humains. Le MCD est organisé autour de trois sous-modèles complémentaires, connectés par la classe commune « NewsItem », représentant une information (voir Figure 1). Le premier concerne le contenu informationnel et permet de représenter les messages diffusés en intégrant leur type, leur narration, les acteurs cités, les émotions mobilisées et leur contexte spatio-temporel. Le second sous-modèle décrit les comportements de diffusion**,** en modélisant les comptes, leurs publications, leurs interactions, les groupes idéologiques et les canaux médiatiques utilisés. Enfin, le troisième sous-modèle introduit la dimension stratégique**,** en structurant les campagnes selon des niveaux hiérarchiques (campagne, opération, action tactique) associés à des objectifs, des contextes et des cibles. Cette structuration permet non seulement de relier contenu et comportement, mais aussi d'inscrire ces éléments dans une logique stratégique explicite, facilitant ainsi l'analyse globale et l'explicabilité des résultats.

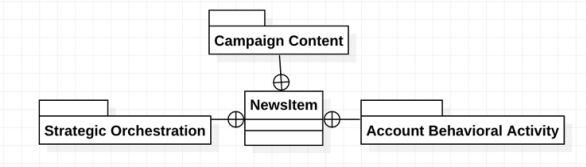


Figure 1- Vue globale des trois sous-modèles connectés

Nous illustrons l'intérêt du modèle en l'appliquant à des articles publiés lors de la campagne de désinformation « Reliable Recent News » (VIGINUM, 2023), attribuée à un réseau pro-russe. En structurant les données extraites de chaque article à l'aide du MCD, il devient possible de relier des articles apparemment indépendants, publiés sur différents sites, à des schémas de diffusion communs et à des objectifs stratégiques

partagés. Le modèle met en évidence des parallèles narratifs, des choix émotionnels récurrents, ainsi que des stratégies de diffusion ciblées, révélant une orchestration cohérente visant à discréditer le soutien occidental à l'Ukraine. Cette étude de cas montre que le MCD permet d'agréger des signaux faibles dispersés, de transformer des observations isolées en une interprétation globale et de justifier analytiquement l'identification d'une campagne coordonnée. Des travaux en cours visent à automatiser le processus de remplissage d'une base de données construite sur le MCD.

Notre approche met en évidence l'intérêt de la modélisation conceptuelle pour dépasser les limites actuelles des outils d'analyse de la désinformation. En fournissant un cadre unifié, explicite et interopérable, le modèle proposé facilite la compréhension humaine des campagnes et ouvre la voie à des systèmes hybrides combinant apprentissage automatique et raisonnement symbolique. Les perspectives incluent l'intégration du MCD dans des architectures d'IA explicables et la constitution de référentiels de campagnes annotées, afin d'améliorer durablement la détection et l'analyse des opérations informationnelles.

Bibliographie

Chalehchaleh, R., Salehi, M., Farahbakhsh, R., Crespi, N.: Brag: a hybrid multifeature framework for fake news detection on social media. Social Network Analysis and Mining 14(1), 35 (2024)

Hu, L., Wei, S., Zhao, Z., Wu, B.: Deep learning for fake news detection: A comprehensive survey. AI open 3, 133–155 (2022)

Lakzaei, B., Haghir Chehreghani, M., Bagheri, A.: Disinformation detection using graph neural networks: a survey. Artificial Intelligence Review 57(3), 52 (2024)

Phan, H.T., Nguyen, N.T., Hwang, D.: Fake news detection: A survey of graph neural network methods. Applied Soft Computing p. 110235 (2023)

Starbird, K., Arif, A., Wilson, T.: Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. Proceedings of the ACM on Human-Computer Interaction 3(CSCW), 1–26 (2019)

Terp, S.J., Breuer, P.: Disarm: a framework for analysis of disinformation campaigns. In: 2022 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA). pp. 1–8. IEEE (2022)

Vargas, L., Emami, P., Traynor, P.: On the detection of disinformation campaign activity with network analysis. In: Proceedings of the 2020 ACM SIGSAC conference on cloud computing security workshop. pp. 133–146 (2020)

VIGINUM: RRN: A complex and persistent information manipulation campaign. https://www.sgdsn.gouv.fr/files/files/Publications/20230719_NP_VIGINUM_RAPPORT-CAMPAGNE-RRN_EN.pdf (July 2023), accessed: 2025-08-13

Zhou, X., Zafarani, R.: A survey of fake news: Fundamental theories, detection methods, and opportunities. ACM Computing Surveys (CSUR) 53(5), 1–40 (2020)

TRACE4PM : Analyse et regroupement des traces pour la modélisation des processus d'interactions utilisateurs dans les systèmes d'information

Marwa Trabelsi[1], Noura Joudieh[2], Amira Ania Dahache[2], Cyrille Suire[2], Ronan Champagnat[2]

1. Centre Génie Industriel
IMT Mines Albi
prenom.nom@mines-albi.fr

2. Laboratoire Informatique, Image et Interaction (L3i)
La Rochelle Université
prenom.nom@univ-lr.fr

RESUME. *Cet article est une synthèse de l'article : Marwa Trabelsi, Noura Joudieh, Amira Ania Dahache, Cyrille Suire, Ronan Champagnat : TRACE4PM: Trace Related Analysis and ClustEring for Process Modeling of Users' Interactions in Information Systems, Proceedings of the International Conference on Research Challenges in Information Science, pp. 3–19, Springer, DOI:10.1007/978-3-031-92474-3_1, 2025.*

MOTS-CLÉS : *API, Modélisation des comportements utilisateurs, Fouille des processus, Clustering de traces, Moodle, Scénarios d'apprentissage*

1. Introduction

L'analyse des comportements utilisateurs dans les systèmes d'information est devenue essentielle pour en améliorer l'ergonomie et l'efficacité. Bien que le Web Usage Mining et le Web Analytics aient contribué à la compréhension des parcours

utilisateurs, ces approches montrent leurs limites face à la complexité croissante des interactions. Le Process Mining (PM) (Van der Aalst, 2016) offre une vision globale des comportements à partir des journaux d'événements, mais génère souvent des modèles complexes nécessitant un regroupement des traces. Pour pallier ces limites, nous proposons TRACE4PM, une API automatisant l'ensemble de la chaîne de traitement, de la préparation des données, du regroupement à la découverte de processus, afin de faciliter l'analyse des comportements utilisateurs, illustrée ici dans le contexte de l'apprentissage en ligne.

2. TRACE4PM

TRACE4PM [1] est une API automatisée de bout en bout dédiée à la modélisation des comportements utilisateurs dans les systèmes d'information via le PM. Conçue selon une architecture orientée micro-services, elle vise à automatiser l'ensemble de la chaîne de traitement habituellement réalisée manuellement dans les travaux de recherche, depuis l'analyse de journaux bruts jusqu'à la découverte et l'évaluation de modèles de processus. L'API s'appuie sur plusieurs micro-services complémentaires, chacun jouant un rôle clé, comme illustré dans la Figure 1. Le **Parser** transforme les journaux d'événements bruts et hétérogènes en journaux structurés, exploitables pour le Process Mining. Le **Tagger**, optionnel et dépendant du domaine, enrichit ces données en traduisant des événements de bas niveau en activités sémantiquement interprétables, facilitant ainsi l'analyse par des utilisateurs non techniques. Le service **Trace Clustering** regroupe les parcours utilisateurs similaires afin de réduire la complexité des modèles, en s'appuyant sur des approches basées sur les traces ou sur des représentations vectorielles (Trabelsi *et al.*, 2021; Joudieh *et al.*, 2024). Enfin, le service de **Process Discovery** génère automatiquement des modèles de processus à l'aide d'algorithmes reconnus, tout en évaluant leur qualité à l'aide de métriques standard.

3. TRACE4PM en Action : Développement d'un plugin Moodle

L'applicabilité de TRACE4PM a été testée à travers un plugin Moodle, illustrant son intégration dans un système d'information réel. Le plugin permet de charger et d'analyser automatiquement les journaux Moodle en exploitant les services de parsing, de clustering de traces et de découverte de processus. À partir de données issues d'un cours universitaire, TRACE4PM a regroupé des parcours d'apprentissage similaires et généré des modèles représentant les comportements typiques des étudiants. L'évaluation menée auprès de chercheurs et d'enseignants a confirmé l'efficacité, la flexibilité et l'utilité de la solution. Les retours mettent en avant l'intérêt de ces modèles pour améliorer les stratégies pédagogiques et identifier les étudiants nécessitant un accompagnement spécifique, tout en soulignant des axes d'amélioration concernant l'ergonomie et l'interprétabilité des résultats.

1. Disponible sur https://github.com/TRACE4PM

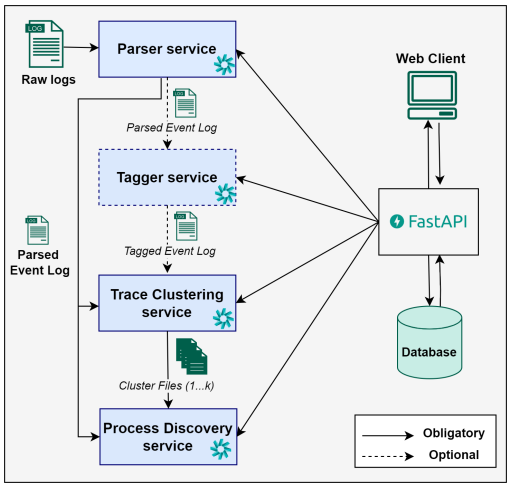


FIGURE 1. *Architecture de l'API TRACE4PM*

4. Conclusion

En conclusion, TRACE4PM est une API automatisée et extensible dédiée à la modélisation des comportements utilisateurs à partir des traces de navigation, en s'appuyant sur les techniques de Process Mining. En intégrant le parsing, l'enrichissement, le clustering de traces et la découverte de processus au sein d'un même cadre, TRACE4PM réduit significativement les efforts manuels et facilite des analyses reproductibles. Son applicabilité a été illustrée par un plugin Moodle, dont l'efficacité a été validée par des retours positifs de chercheurs et d'enseignants. Les perspectives futures incluent l'enrichissement des algorithmes intégrés, l'extension de l'évaluation à d'autres contextes et l'amélioration de l'explicabilité des résultats afin de rendre l'outil plus accessible aux utilisateurs non experts.

Bibliographie

Joudieh N., Trabelsi M., Champagnat R., Rabah M., Eteokleous N., « Using Trace Clustering to Group Learning Scenarios : An Adaptation of FSS-Encoding to Moodle Logs Use Case. », *CSEDU (2)*, p. 247-254, 2024.

Trabelsi M., Suire C., Morcos J., Champagnat R., « A New Methodology to Bring Out Typical Users Interactions in Digital Libraries », *2021 ACM/IEEE Joint Conf. on Digital Libraries (JCDL)*, IEEE, p. 11-20, 2021.

Van der Aalst W., *Process mining : data science in action*, Springer, 2016.

Fouille de personas via fouille de processus et apprentissage automatique non-supervisé

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. (2025). Persona Mining Using Process Mining and Unsupervised Machine-Learning, KES 2025, Osaka Japan.

1. Introduction

Dans un contexte de transformation numérique généralisée, les systèmes d'information et plateformes numériques sont omniprésents et soutiennent des domaines variés tels que l'éducation, la santé, le commerce électronique ou encore les écosystèmes organisationnels. Malgré cette omniprésence, l'expérience utilisateur (User eXperience – UX) demeure souvent insatisfaisante en raison de conceptions peu intuitives, de données incomplètes ou redondantes, et de difficultés d'accès à l'information pertinente. Ces limitations peuvent affecter négativement l'engagement, la satisfaction et la fidélité des utilisateurs (Sauro, Lewis, 2016).

L'amélioration de l'UX repose en grande partie sur la capacité des systèmes à proposer des recommandations personnalisées, adaptées aux besoins, aux intentions et au contexte des utilisateurs (Fernández-Portillo *et al*., 2024). Dans cette perspective, la notion de persona joue un rôle central. Un persona représente un type d'utilisateur caractérisé par des comportements, des objectifs et des attentes similaires. Traditionnellement, les personas sont construits à partir d'enquêtes, d'entretiens ou de questionnaires, méthodes qui sont coûteuses, intrusives et parfois biaisées.

Nous partons du constat que les traces d'utilisation laissées par les utilisateurs dans les systèmes, sous forme de journaux d'événement (event logs), constituent une source de données riche, objective et facilement accessible pour analyser les comportements réels. Cependant, la littérature existante en fouille de comportement (behavior mining) ne propose pas de méthodes permettant de construire des personas directement à partir de ces event logs. Ainsi, l'objectif principal de l'article est de proposer une approche innovante pour identifier et construire automatiquement des personas à partir des traces d'activité des utilisateurs, en combinant le process mining et des techniques d'apprentissage automatique non supervisé.

2. Fouille de personas basée sur l'apprentissage non-supervisé

L'approche proposée vise à extraire des personas à partir de journaux d'événements générés par les interactions des utilisateurs avec un système. Les event logs

doivent contenir au minimum trois attributs essentiels: 1) un identifiant de cas (USE CASE ID), 2) une activité (ACTIVITY) et 3) un horodatage (TIMESTAMP). Quelle que soit la forme initiale des données (CSV, XES, TXT), celles-ci sont prétraitées et normalisées sous forme de DataFrame afin de faciliter les étapes suivantes.

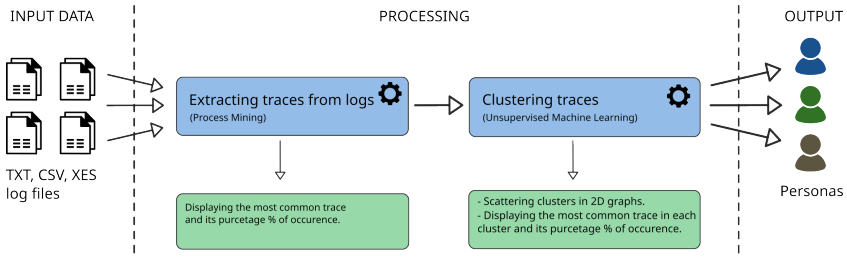


FIGURE 1 – Vue globale de la méthode proposée

Comme le montre la figure 1, à partir des event logs, nous appliquons des techniques de process mining pour extraire des traces, c'est-à-dire des séquences chronologiques d'activités réalisées par un utilisateur au cours d'un cas donné. Chaque trace représente un comportement utilisateur complet, depuis le début jusqu'à la fin d'une interaction ou d'une session. Ces traces peuvent être représentées sous forme de graphes dirigés, reflétant l'ordre temporel des événements. Nous observons une grande diversité de traces, allant de séquences simples à des enchaînements complexes, avec des répétitions d'activités possibles. Une première analyse consiste à identifier les traces les plus fréquentes et leur taux d'occurrence dans le jeu de données.

La seconde étape consiste à regrouper les traces similaires afin d'identifier des comportements proches. Pour cela, nous utilisons le clustering non supervisé, en particulier l'algorithme k-means, appliqué aux représentations des traces. Chaque cluster correspond alors à un ensemble de traces similaires, interprété comme une classe de comportement ou un persona.

Afin de déterminer le nombre optimal de clusters, et donc de personas, deux métriques de validation sont utilisées: le Silhouette Score (Shahapure, Nicholas, 2020) et le Davies-Bouldin Index (Xiao *et al.*, 2017). Ces métriques permettent d'évaluer respectivement la cohésion interne des clusters et leur séparation. L'analyse de ces indicateurs guide le choix du nombre de personas pertinents pour chaque jeu de données. Une fois les clusters établis, chaque utilisateur est associé à un persona, ce qui permet de relier les comportements observés à des profils d'utilisateurs exploitables dans des systèmes de recommandation.

3. Évaluation expérimentale

L'approche est évaluée sur trois jeux de données, dont deux sont décrits en détail dans l'article : 1) Permit Log Dataset: issu du BPI Challenge 2020 (ICPM), ce jeu de données concerne des demandes d'autorisation de déplacement dans une organisation.

Il contient plus de 86 000 événements et présente une forte variabilité des comportements utilisateurs. 2) Call Center Dataset : provenant des logs de démonstration de Fluxicon Disco, ce dataset décrit les interactions entre clients et un centre d'appel, incluant redirections, rappels et clôtures d'appels. Ces deux jeux de données respectent les prérequis définis pour l'extraction des traces et permettent de tester la robustesse de l'approche dans des contextes différents.

Les résultats montrent qu'il est possible d'identifier des clusters stables et interprétables à partir des traces extraites. Pour le dataset Permit Log, les métriques Silhouette et Davies-Bouldin indiquent que trois clusters offrent le meilleur compromis entre cohésion et séparation. Pour le dataset Call Center, les résultats sont plus nuancés, mais suggèrent que quatre clusters constituent un choix pertinent.

Les visualisations en deux dimensions confirment l'existence de groupes bien distincts, chacun correspondant à un type de comportement utilisateur spécifique. L'analyse qualitative des traces dominantes dans chaque cluster révèle des différences significatives dans les parcours utilisateurs, confirmant que chaque cluster peut être interprété comme un persona distinct.

4. Conclusion

L'étude propose une approche innovante et non intrusive pour le mining de personas basée sur l'analyse de traces d'utilisation réelles. En combinant process mining et apprentissage non supervisé, nous démontrons qu'il est possible de construire automatiquement des personas fiables à partir d'event logs, sans recourir à des enquêtes ou questionnaires utilisateurs. Cette contribution ouvre des perspectives importantes pour les systèmes de recommandation et l'orchestration d'entités dans les écosystèmes numériques, en intégrant le comportement réel des utilisateurs dans les processus de personnalisation.

Les travaux futurs envisagent d'enrichir cette approche par des données qualitatives (enquêtes, interviews), d'expérimenter d'autres algorithmes de clustering (DBSCAN, clustering hiérarchique, modèles de mélanges gaussiens) et d'évaluer l'impact concret des personas sur la performance des recommandations dans des écosystèmes numériques réels.

Bibliographie

Fernández-Portillo A., Ramos-Vecino N., Ramos-Mariño A., Cachón-Rodríguez G. (2024). How the digital business ecosystem affects stakeholder satisfaction: its impact on business performance. *Review of Managerial Science*, p. 1–20.

Sauro J., Lewis J. R. (2016). *Quantifying the user experience: Practical statistics for user research*. Morgan Kaufmann.

Shahapure K. R., Nicholas C. (2020). Cluster quality analysis using silhouette score. , vol. 1, n° 1, p. 747-748.

Xiao J., Lu J., Li X. (2017). Davies bouldin index based hierarchical initialization k-means. *Intelligent Data Analysis*, vol. 21, n° 6, p. 1327–1338.

Piloter la Durabilité par la Fouille de Processus

Cadres d'Analyse et Indicateurs pour l'Aide à la Décision dans les Systèmes d'Information

Nikita Valenza, Rébecca Deneckère

Centre de Recherche en Informatique
Université Paris 1 Panthéon-Sorbonne, France
nikita.valenza@yahoo.com, rebecca.deneckere@univ-paris1.fr

RESUME. La durabilité est un enjeu central pour les systèmes d'information décisionnels, en particulier dans des contextes organisationnels et industriels soumis à de fortes contraintes environnementales, économiques et sociales. La fouille de processus s'impose comme une approche prometteuse pour relier l'exécution réelle des processus aux objectifs de durabilité, à partir de l'analyse de traces issues des systèmes d'information. Cet article propose une revue analytique structurée des travaux mobilisant la fouille de processus pour l'évaluation et le pilotage de la durabilité. Les résultats mettent en évidence une forte hétérogénéité des approches, tant sur les sources de données que sur les cadres conceptuels, ainsi qu'un déséquilibre marqué entre les dimensions environnementales et sociales. L'article discute les implications de ces constats pour la conception de systèmes d'information décisionnels durables et identifie plusieurs pistes de recherche, notamment autour de l'intégration d'indicateurs éthiques et de l'actionnabilité des résultats.

ABSTRACT. Sustainability is a key concern for decision-support information systems, particularly in organizational and industrial contexts facing strong environmental, economic, and social constraints. Process Mining emerges as a promising approach to connect actual process execution with sustainability objectives through the analysis of event logs extracted from information systems. This paper presents a structured analytical review of research works applying Process Mining to sustainability assessment and decision support. The results highlight a strong diversity of approaches in terms of data sources and conceptual frameworks, as well as a marked imbalance in favor of environmental dimensions compared to social aspects. The paper discusses the implications of these findings for the design of sustainable decision-support information systems and outlines several research perspectives, particularly related to ethical indicators and the actionability of Process Mining outcomes.

MOTS-CLES : Fouille de processus, Durabilité, Systèmes d'Information Décisionnels, Indicateurs de Performance, Aide à la Décision.

KEYWORDS: Process Mining, Sustainability, Decision-support Information Systems, Performance Indicators, Decision-making

1. Introduction

La durabilité constitue un enjeu structurant pour les organisations, impliquant l'intégration conjointe de dimensions environnementales, économiques et sociales dans les systèmes d'information décisionnels (Rabbanee *et al.,* 2023), (Union Européenne, 2024). Dans ce contexte, la fouille de processus permet d'exploiter les traces issues des systèmes d'information afin d'analyser l'exécution réelle des processus et de soutenir la prise de décision.

Plusieurs travaux récents mettent en évidence le potentiel de la fouille de processus pour analyser des dimensions environnementales de la durabilité, notamment la consommation d'énergie, les émissions de gaz à effet de serre ou l'optimisation des flux de production (Watanabe *et al.,* 2017; Rai et Daniels, 2016; Kroeger *et al.,* 2024). D'autres approches étendent cette perspective à des cadres d'évaluation plus larges, intégrant des dimensions économiques et sociales afin d'évaluer la durabilité globale de systèmes productifs ou de processus métiers dans une logique de Triple Bilan (Triple Bottom Line) (Kroeger *et al.,* 2024; Docekalova, 2017). Toutefois, malgré cet intérêt croissant, la littérature demeure hétérogène et fragmentée. Les travaux diffèrent fortement selon les types de données mobilisées, les cadres de modélisation retenus, les méthodes de construction des indicateurs et les niveaux d'analyse considérés, qu'ils soient organisationnels, industriels ou intégrés. Cette diversité rend difficile l'identification de cadres analytiques cohérents permettant de guider la conception de SI décisionnels durables. En outre, la majorité des travaux se concentre sur les dimensions environnementales et économiques, tandis que les enjeux sociaux et éthiques liés à l'exploitation des données de processus, tels que l'équité[1], la confidentialité ou la transparence des analyses, restent encore marginalement abordés (van der Aalst, 2016; Park et van der Aalst, 2022).

Ces constats soulèvent une question centrale pour la recherche en systèmes d'information : comment la fouille de processus peut-elle être mobilisée de manière structurée pour soutenir l'aide à la décision en matière de durabilité au sein des SI ? Cette question dépasse la simple application de techniques analytiques à des indicateurs environnementaux. Elle interroge la manière dont les données de processus sont constituées, modélisées et interprétées, ainsi que les cadres conceptuels et décisionnels dans lesquels elles s'inscrivent, afin de produire des résultats exploitables, responsables et alignés avec les objectifs de durabilité. Pour répondre à cette question, cet article propose une revue analytique structurée des travaux mobilisant la fouille de processus au service de la durabilité et de l'aide à la décision. Contrairement aux revues existantes majoritairement orientées vers des cadres conceptuels ou sectoriels, ce travail adopte une grille de lecture explicitement ancrée dans les systèmes d'information décisionnels, en analysant conjointement la constitution des données, les cadres d'évaluation et les mécanismes de traduction des

[1] Dans la littérature anglo-saxonne en fouille de processus et en science des données, le terme *fairness* désigne des propriétés formalisables liées à l'absence de biais et de discrimination dans les données, les modèles ou les résultats algorithmiques ; il est ici rapproché du terme français *équité*, entendu au sens opérationnel et non uniquement normatif.

résultats analytiques en décisions. L'objectif est d'identifier les choix structurants et les angles morts de la littérature afin de guider la conception de SI décisionnels durables.

L'article est structuré comme suit. La section 2 décrit quelques fondamentaux et la section 3 les travaux connexes. La méthodologie de recherche est expliquée en section 4. La section 5 propose les résultats de l'analyse organisée selon plusieurs axes. Nous proposons une discussion en section 6 et nous concluons en section 7.

2. Fondamentaux

Cette section présente les fondements théoriques mobilisés dans cet article et introduit le cadre qui structure l'analyse des travaux étudiés. Elle vise à clarifier les notions de fouille de processus, de durabilité et de SI décisionnels, ainsi que leurs articulations, afin de fournir une base cohérente pour l'analyse comparative.

Fouille de processus et systèmes d'information. La fouille de processus désigne un ensemble de techniques visant à analyser l'exécution réelle des processus à partir des traces enregistrées. Ces traces associent des informations telles qu'un identifiant de cas, une activité, un horodatage et, selon les contextes, des attributs complémentaires liés aux ressources ou aux objets manipulés (Bauer, 2024; van Dongen et van der Aalst, 2009). À la différence des approches traditionnelles de modélisation des processus, la fouille ne repose pas sur des modèles prescriptifs, mais sur l'observation empirique des comportements effectifs des systèmes. Les ERP, MES ou systèmes de production automatisés génèrent des volumes importants de données événementielles, qui peuvent être exploitées pour reconstruire des modèles de processus, analyser des performances ou vérifier la conformité des pratiques (Bauer, 2024). La fouille de processus est donc un moyen de transformer des données opérationnelles en connaissances exploitables pour la prise de décision. Les principaux types de techniques de fouille de processus incluent la découverte de processus, l'analyse de conformité et l'analyse de performance (van Dongen et van der Aalst, 2009). La découverte de processus vise à produire des modèles représentant les enchaînements d'activités observés, souvent sous forme de réseaux de Petri ou de variantes dérivées. L'analyse de conformité permet de comparer ces modèles aux processus prescrits ou attendus, tandis que l'analyse de performance exploite les dimensions temporelles et quantitatives des traces afin d'identifier des inefficiences ou des goulots d'étranglement. Ces techniques constituent le socle analytique sur lequel reposent les applications de la fouille de processus à des objectifs de durabilité.

Durabilité et SI décisionnels. La durabilité est généralement appréhendée à travers la notion de Triple Bilan (United Nations, 1992; Rabbanee *et al.,* 2023). Pour les systèmes d'information, cela se traduit par la nécessité de concevoir des outils capables de mesurer, d'analyser et de piloter ces différentes dimensions de manière intégrée. Les SI décisionnels jouent ici un rôle clé, en agrégeant des données hétérogènes et en produisant des indicateurs destinés à soutenir les choix stratégiques et opérationnels des organisations. Les indicateurs de durabilité, souvent formalisés sous forme de KPI, constituent un élément central de ces systèmes. Ils permettent de traduire des objectifs abstraits, tels que la réduction de l'empreinte carbone ou

l'amélioration des conditions de travail, en mesures opérationnelles exploitables (Docekalova, 2017). Toutefois, la littérature souligne la difficulté de construire des systèmes d'indicateurs cohérents, comparables et adaptés aux contextes organisationnels, en raison de la diversité des sources de données, des niveaux d'analyse et des référentiels mobilisés (Docekalova, 2017; United Nations, 2023). Dans ce contexte, les SI décisionnels orientés durabilité doivent répondre à plusieurs exigences. Ils doivent être capables d'intégrer des données issues de processus opérationnels, de supporter des analyses multicritères et de produire des résultats interprétables par les décideurs. Ils doivent également prendre en compte des contraintes éthiques et réglementaires, notamment en matière de confidentialité et de transparence, afin de garantir la légitimité des décisions produites (van der Aalst, 2016).

Articulation entre fouille de processus et durabilité. L'application de la fouille de processus à la durabilité repose sur l'hypothèse selon laquelle les impacts environnementaux, économiques et sociaux des organisations sont en grande partie déterminés par l'exécution concrète de leurs processus. En analysant les processus tels qu'ils sont réellement réalisés, la fouille de processus offre la possibilité de relier des indicateurs de durabilité à des activités, des enchaînements ou des configurations organisationnelles spécifiques (Graves et van der Aalst, 2023; Kroeger *et al.,* 2024). La littérature montre que cette articulation peut prendre des formes variées. Dans certains cas, la fouille est utilisée pour enrichir des analyses existantes en y intégrant des indicateurs environnementaux ou économiques (Watanabe *et al.,* 2017; Rai et Daniels, 2016). Dans d'autres cas, il constitue le cœur d'un cadre analytique plus large, combinant modélisation des processus, simulation et construction de systèmes d'indicateurs orientés durabilité (Kroeger *et al.,* 2024; Gribaudo et Manini, 2020). Cependant, la construction de traces adaptées à l'analyse de la durabilité nécessite souvent de combiner des données hétérogènes, issues à la fois de SI métier et de sources externes, telles que des bases de données environnementales ou des référentiels normatifs. De plus, le choix des indicateurs et des cadres d'évaluation influence fortement les résultats produits et leur interprétation par les décideurs, ce qui renforce la nécessité d'une approche structurée et explicite.

Cadre de travail de l'analyse. Ce travail d'analyse adopte un cadre articulant trois niveaux complémentaires, correspondant aux principales décisions de conception d'un SI décisionnel orienté durabilité. Le premier niveau porte sur les données et la constitution des traces, conditionnant la mesurabilité des impacts. Dans ce travail, les systèmes d'information industriels (MES, CPS, systèmes de capteurs) sont considérés comme des SI à part entière, opérant à un niveau opérationnel et cyber-physique, et complémentaires des systèmes d'information décisionnels de l'entreprise. Le deuxième concerne les cadres de modélisation et d'évaluation, qui structurent l'interprétation des processus. Le troisième porte sur les systèmes d'indicateurs et l'aide à la décision, où se matérialise la valeur décisionnelle des analyses. Ce cadre permet ainsi d'analyser les travaux non seulement selon leurs résultats, mais selon les choix de conception qu'ils impliquent pour les SI. La Figure 1 illustre le rôle de la fouille comme médiateur entre les SI et l'aide à la décision orientée durabilité, en mettant en évidence les niveaux de données, d'analyse et de décision, ainsi que le caractère transversal des enjeux sociaux et éthiques.

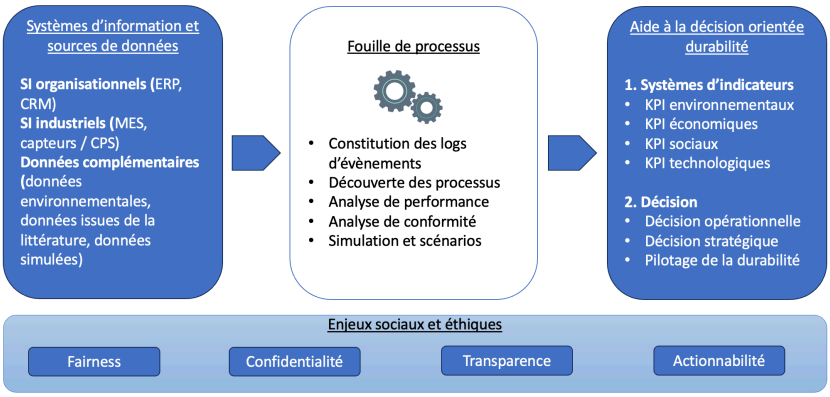


Figure 1 – Cadre de la fouille de processus au service de la durabilité dans les SI.

Cette figure met en évidence l'articulation entre les sources de données, les analyses de processus et la construction de systèmes d'indicateurs, ainsi que le caractère transversal des enjeux sociaux et éthiques.

Afin d'illustrer concrètement l'articulation entre données de processus, indicateurs de durabilité et aide à la décision, considérons un processus de production industrielle. Des données issues de capteurs peuvent être associées aux activités du processus, par exemple en termes de consommation énergétique. Après constitution des traces et enrichissement sémantique, la fouille de processus permet d'identifier différentes variantes d'exécution. L'analyse peut alors révéler que certaines variantes sont plus énergivores que d'autres, en raison de séquences d'activités ou de temps d'attente spécifiques. Cette capacité à relier directement des indicateurs de durabilité à des configurations réelles de processus illustre le rôle de la fouille comme médiateur entre données opérationnelles et décision, en identifiant des leviers d'optimisation ciblés.

3. Travaux Connexes

Les travaux mobilisant la fouille de processus dans une perspective de durabilité s'inscrivent à l'intersection de plusieurs champs de recherche, notamment le Business Process Management, les SI décisionnels et les études sur la durabilité.

Application de la fouille de processus pour l'analyse et l'optimisation des performances opérationnelles. Ces travaux ont structuré les techniques de découverte, de conformité et de performance à partir de logs d'événements (Bauer, 2024; van Dongen et van der Aalst, 2009). Bien que principalement orientées vers l'efficacité, la réduction des coûts ou la conformité réglementaire, ces approches constituent le socle méthodologique sur lequel se sont appuyées les applications ultérieures de la fouille à des problématiques environnementales et sociétales.

Intégration des préoccupations environnementales dans l'analyse des processus. Ces travaux mobilisent la fouille de processus pour analyser la consommation

énergétique, les émissions de gaz à effet de serre ou l'utilisation des ressources dans des contextes industriels et de production (Watanabe *et al.,* 2017; Rai et Daniels, 2016). Toutefois, ces approches restent souvent focalisées sur une dimension spécifique de la durabilité et ne proposent pas de cadre décisionnel intégrant l'ensemble des dimensions possibles.

Vision intégrée de la durabilité. Ces travaux mobilisent des cadres tels que le Triple Bilan qui renvoie à l'évaluation conjointe des différentes dimensions. Certaines contributions proposent des systèmes d'indicateurs combinant fouille de processus, simulation et méthodes d'agrégation afin d'évaluer la durabilité de systèmes productifs ou de réseaux de production (Kroeger *et al.,* 2024; Gribaudo et Manini, 2020). D'autres s'appuient sur des techniques telles que la logique floue pour traiter des indicateurs qualitatifs ou imprécis et comparer la performance durable de processus ou d'organisations (Docekalova, 2017; Safitri *et al.,* 2018). Ces approches élargissent le périmètre de l'analyse, mais restent hétérogènes dans leurs choix méthodologiques. Un travail de référence (Graves et van der Aalst, 2023) offre une vue d'ensemble des applications de la fouille à la durabilité et propose un cadre, le PM4S, visant à relier la fouille aux principes de l'économie circulaire. L'accent est mis sur la gestion des flux de ressources, la logistique inverse et l'usage de modèles orientés objets. Si ce travail fournit une base théorique solide et met en lumière le potentiel de la fouille pour soutenir des pratiques durables, il adopte principalement une perspective conceptuelle et sectorielle, centrée sur l'économie circulaire, sans proposer une analyse détaillée des méthodes de constitution des données, des cadres d'évaluation ou des systèmes d'indicateurs mobilisés dans les travaux étudiés.

Par ailleurs, les enjeux sociaux et éthiques associés à l'exploitation des données de processus sont encore peu présents dans les travaux existants, excepté (van der Aalst, 2016) qui aborde explicitement des questions d'équité, de confidentialité et de transparence dans l'analyse des données. De même, les travaux sur l'Action-Oriented Process Mining (Park et van der Aalst, 2022) mettent en évidence l'importance de l'actionnabilité des résultats et du passage des analyses à des décisions concrètes. Néanmoins, ces contributions restent rarement mobilisées de manière systématique dans les travaux appliquant la fouille de processus à la durabilité. La littérature met en évidence le potentiel de la fouille de processus pour soutenir des objectifs de durabilité, mais elle demeure fragmentée selon les dimensions de la durabilité considérées, les types de données exploitées et les cadres analytiques retenus. Peu de travaux proposent une grille de lecture transversale permettant d'analyser conjointement les méthodes de constitution des données, les cadres de modélisation, les systèmes d'indicateurs et les enjeux sociaux associés, dans une perspective explicitement orientée SI décisionnels. C'est précisément cette lacune que le présent travail se propose de combler, en offrant une revue structurée des approches et en mettant en évidence leurs implications pour la conception de SI durables.

4. Méthode de Recherche

Cet article adopte une démarche de revue analytique structurée visant à analyser de manière systématique les travaux mobilisant la fouille de processus pour soutenir

l'aide à la décision en matière de durabilité dans les systèmes d'information. La méthodologie repose sur une sélection de contributions scientifiques et sur une grille d'analyse construite à partir du cadre de travail présenté précédemment.

Définition de la question de recherche. La question de recherche principale, formulée dans l'introduction, est la suivante : **Comment la fouille de processus peut-elle être mobilisée de manière structurée pour soutenir l'aide à la décision en matière de durabilité au sein des systèmes d'information ?** Cette question est déclinée en quatre sous-questions de recherche, qui permettent d'examiner successivement les choix méthodologiques, conceptuels et décisionnels des approches étudiées : *RQ1 : Quelles sont les méthodes de constitution des données et des traces mobilisées pour la quantification d'indicateurs de durabilité à l'aide de la fouille de processus ? RQ2 : Quels cadres de modélisation et d'évaluation sont proposés pour intégrer la durabilité dans l'analyse des processus ? RQ3 : Comment les différentes perspectives de la durabilité influencent-elles la construction et l'usage des systèmes d'indicateurs pour l'aide à la décision ? RQ4 : Quels sont les enjeux sociaux et éthiques associés à l'application de la fouille de processus à la durabilité dans des environnements de données à grande échelle ?*

Sélection du corpus. La sélection du corpus s'appuie sur une recherche bibliographique structurée, réalisée sur la base de données Scopus. Les mots-clés ont été définis de manière à couvrir trois axes principaux : la fouille de processus, les indicateurs et métriques, ainsi que la durabilité. Ces axes ont été combinés à l'aide d'opérateurs logiques afin d'identifier des travaux traitant explicitement de l'articulation entre fouille et durabilité. La chaîne de recherche utilisée est la suivante: (TITLE-ABS-KEY("Process Mining") OR TITLE-ABS-KEY("Event Log") OR TITLE-ABS-KEY("Process Optimization") OR TITLE-ABS-KEY("Petri Nets")) AND (TITLE-ABS-KEY("Metrics") OR TITLE-ABS-KEY("Data Analytics") OR TITLE-ABS-KEY("Indicators") OR TITLE-ABS-KEY("Measurements") OR TITLE-ABS-KEY("Measure")) AND (TITLE-ABS-KEY("Sustainable Development Goals") OR TITLE-ABS-KEY("Sustainability") OR TITLE-ABS-KEY("Climate Action") OR TITLE-ABS-KEY("Corporate Social Responsibility") OR TITLE-ABS-KEY("Circular Economy")) AND PUBYEAR > 2015. Cette chaîne a permis d'identifier 161 sources possibles. La base Scopus a été retenue pour sa couverture large et sa structuration des métadonnées. L'absence d'intégration d'autres bases constitue une limite, mais ne remet pas en cause la diversité des approches identifiées.

Les critères d'inclusion sont les suivants : (i) articles scientifiques évalués par les pairs ; (ii) en anglais ou en français ; (iii) publiés après 2015, pour garantir la prise en compte des évolutions récentes de la fouille de processus et des enjeux de durabilité ; (iv) contributions mobilisant explicitement des techniques de fouille ou des approches assimilées à partir de traces ; (v) présence explicite d'indicateurs ou de métriques liés à la durabilité. Les critères d'exclusion comprennent les travaux portant sur l'extraction minière au sens géologique, les articles trop éloignés du champ des SI, ou ceux ne proposant pas d'analyse exploitable du point de vue des processus. Cette démarche a conduit à l'identification de 10 articles constituant le corpus analysé. Ce choix reflète un compromis entre exhaustivité et profondeur analytique, l'objectif de la revue n'étant pas de couvrir l'ensemble des publications liées à la durabilité, mais d'analyser en détail des travaux proposant une articulation explicite entre fouille de

processus, indicateurs et aide à la décision. La sélection privilégie la diversité des cadres analytiques et des niveaux d'analyse.

La démarche s'inspire des principes des revues de type PRISMA. Toutefois, elle ne vise pas l'exhaustivité statistique mais une analyse qualitative structurée. Certaines références non académiques sont mobilisées de manière complémentaire sans constituer le cœur du corpus analysé.

Démarche d'analyse. L'analyse repose sur une grille construite à partir du cadre présenté en section 2. Chaque contribution est analysée selon quatre axes correspondant aux sous-questions de recherche : les sources de données et les méthodes de constitution des traces, les cadres de modélisation et d'évaluation mobilisés, les systèmes d'indicateurs et leur rôle dans l'aide à la décision et les enjeux sociaux et éthiques abordés. Cela permet une comparaison systématique des approches étudiées, en mettant en évidence leurs points communs, leurs divergences et leurs limites. L'objectif n'est pas d'évaluer la performance relative des approches, mais d'identifier des tendances, des choix structurants et des lacunes potentielles. Il s'agit de mettre en évidence les choix de conception sous-jacents pour les SI décisionnels.

5. Analyse

Cette section présente l'analyse du corpus par sous-questions de recherche.

5.1. RQ1 – Méthodes de constitution des données et des traces

La première question de recherche porte sur les méthodes mobilisées pour constituer les données nécessaires à l'application de la fouille de processus dans une perspective de durabilité. L'analyse met en évidence une forte hétérogénéité des sources de données et des niveaux de granularité retenus. Le Tableau 1 propose une typologie des travaux analysés selon les types de données mobilisées et les niveaux d'analyse. Il positionne les travaux mobilisant explicitement des données de processus dans une perspective d'analyse (les contributions à dominante conceptuelle, qui ne se prêtent pas à une classification selon les types de données et les niveaux d'analyse, sont discutées séparément dans les travaux connexes et la discussion).

Une première distinction peut être établie entre les approches reposant principalement sur des données issues de SI organisationnels et celles mobilisant des données de niveau industriel. Les premières exploitent des traces extraites d'ERP ou de systèmes métiers afin d'analyser des processus à un niveau relativement agrégé, souvent orienté vers la gestion des flux, la conformité ou la performance organisationnelle (Docekalova, 2017; Park et van der Aalst, 2022). Ces approches facilitent l'intégration de considérations économiques et sociales, mais elles offrent une vision limitée des impacts environnementaux liés aux opérations physiques.

À l'inverse, les travaux centrés sur des environnements industriels exploitent des données de production à plus forte granularité, issues de MES, de capteurs ou de systèmes cyber-physiques (Watanabe *et al.,* 2017; Rai et Daniels, 2016; Kumbhar *et*

Tableau 1. Typologie des approches de fouille de processus

Type de données mobilisées / Niveaux	Organisationnel	Production / industriel	Intégré
Données organisationnelles de haut niveau logs ERP, données métier, enquêtes, entretiens	(Docekalova, 2017), (Safitri *et al.,* 2018), (Park et van der Aalst, 2022)		
Données opérationnelles de bas niveau MES, capteurs, logs machines, CPS		(Watanabe *et al.,* 2017), (Rai et Daniels, 2016), (Kumbhar *et al.,* 2022)	
Données hybrides / enrichies combinaison SI + données industrielles + littérature + simulation			(Kroeger *et al.,* 2024), (Gribaudo et Manini, 2020), (Acerbi, 2022)

al., 2022). Ces données permettent d'associer des indicateurs environnementaux, tels que la consommation énergétique ou les émissions, à des activités de processus spécifiques. Toutefois, leur intégration dans des traces exploitables pour la fouille de processus nécessite souvent des opérations complexes de prétraitement, de synchronisation et d'enrichissement des données.

Plusieurs travaux proposent des approches hybrides combinant données organisationnelles et industrielles, parfois enrichies par des données issues de la littérature ou de référentiels environnementaux (Kroeger *et al.,* 2024). Cette hybridation apparaît comme nécessaire pour construire des analyses de durabilité plus complètes, mais soulève des défis importants en termes de cohérence des données et de charge de mise en œuvre pour les SI.

Enfin, un nombre significatif de travaux recourt à la simulation, notamment à événements discrets, pour générer des données lorsque les traces réelles sont incomplètes ou indisponibles (Rai et Daniels, 2016; Kroeger *et al.,* 2024; Gribaudo et Manini, 2020). Ces approches permettent d'explorer des scénarios prospectifs.

Préparation des traces pour la fouille de processus. Au-delà des sources de données, l'exploitation de la fouille de processus repose sur des étapes de préparation des traces souvent peu explicitées dans les travaux analysés. Pourtant, ces transformations conditionnent directement la qualité des analyses produites. La construction de logs exploitables implique plusieurs opérations clés. Tout d'abord, un filtrage des données est nécessaire afin d'éliminer les événements incomplets, les doublons ou les traces bruitées. Ensuite, la structuration des cas constitue une étape centrale, consistant à définir ce qui constitue une instance de processus (par exemple une commande, un lot de production ou un cycle machine). Les données issues de sources hétérogènes doivent également être synchronisées temporellement et intégrées dans un format cohérent. Enfin, un enrichissement sémantique est souvent requis pour associer aux événements des attributs liés à la durabilité, tels que la consommation énergétique, les émissions ou l'utilisation de ressources. Ces étapes de transformation traduisent le passage de données brutes issues des SI vers des logs orientés processus, directement exploitables par les techniques de fouille. Leur

absence ou leur simplification limite fortement la capacité à analyser les processus réels et à produire des indicateurs pertinents. En particulier, la définition du cas et le choix du niveau de granularité influencent directement les modèles de processus découverts et, par conséquent, les indicateurs de durabilité qui peuvent être dérivés.

5.2. RQ2 – Cadres de modélisation et d'évaluation de la durabilité

Les cadres analysés peuvent être distingués selon leur capacité à exploiter explicitement la structure des processus, avec trois niveaux possibles : le niveau *activité*, où les indicateurs sont associés à des tâches individuelles, le niveau *trace*, où les analyses portent sur des instances complètes de processus, et le niveau des *variantes*, où les différentes trajectoires d'exécution sont comparées. Cette distinction est essentielle pour caractériser l'apport spécifique de la fouille de processus par rapport à des approches analytiques classiques.

L'analyse révèle également trois grandes familles d'approches.

Cadres orientés pilotage organisationnel. Ils se concentrent sur l'évaluation de processus métiers ou de chaînes de valeur à un niveau agrégé. Ils s'appuient sur des indicateurs issus des SI décisionnels et mobilisent des techniques telles que l'analyse multidimensionnelle ou la logique floue afin de traiter des données qualitatives ou imprécises (Docekalova, 2017; Safitri *et al.,* 2018). Ces approches sont particulièrement adaptées à des comparaisons inter-organisationnelles ou à des évaluations globales de la durabilité, mais elles offrent une visibilité limitée sur les mécanismes opérationnels sous-jacents.

Cadres orientés performance productive. Ils s'intéressent aux processus industriels et aux flux physiques. Ils mobilisent fréquemment des modèles de simulation, des réseaux de Petri ou des approches issues de la physique des systèmes de production pour analyser l'efficacité énergétique, les goulots d'étranglement ou l'utilisation des ressources (Watanabe *et al.,* 2017; Rai et Daniels, 2016; Kumbhar *et al.,* 2022). Ces cadres permettent une analyse fine des impacts environnementaux, mais ils sont souvent moins intégrés aux SI décisionnels de l'organisation.

Cadres systémiques ou intégrateurs. Ils combinent fouille, simulation et systèmes d'indicateurs afin d'évaluer la durabilité à plusieurs niveaux, du poste de travail au réseau de production (Kroeger *et al.,* 2024; Gribaudo et Manini, 2020; Acerbi, 2022). Ces approches apparaissent comme les plus prometteuses pour soutenir l'aide à la décision, mais elles impliquent une complexité technique et organisationnelle élevée, notamment en termes d'intégration des données et de gouvernance des modèles.

5.3. RQ3 – Systèmes d'indicateurs et aide à la décision

Les systèmes d'indicateurs identifiés peuvent être distingués selon leur ancrage dans les processus. Certains indicateurs sont calculés indépendamment de la structure des processus, tandis que d'autres sont explicitement liés aux activités, aux séquences

ou aux variantes d'exécution. Cette distinction permet d'évaluer dans quelle mesure les approches exploitent réellement les spécificités de la fouille de processus.

L'analyse montre que la majorité des travaux s'inscrit dans une logique de Triple Bilan, avec une prédominance marquée des indicateurs environnementaux. Le Tableau 2 synthétise les dimensions de la durabilité abordées dans les travaux analysés, les types d'indicateurs mobilisés et leur rôle dans l'aide à la décision.

Tableau 2 – Dimensions de la durabilité

Dimension de durabilité	**Types d'indicateurs**	**Méthodes de construction**	**Finalité décisionnelle**
Environnementale (Watanabe *et al.,* 2017), (Rai et Daniels, 2016), (Kroeger *et al.,* 2024), (Gribaudo et Manini, 2020), (Acerbi, 2022)	- Consommation énergétique - émissions de GES - flux de matières, déchets - efficacité énergétique	- Mesures directes à partir des logs - couplage fouille–simulation (DES) - normalisation par objectifs	- Optimisation des processus, réduction de l'empreinte environnementale - comparaison de scénarios
Économique (Kroeger *et al.,* 2024), (Kumbhar *et al.,* 2022), (Docekalova, 2017), (Park et van der Aalst, 2022)	- Coûts de production - productivité - taux d'utilisation - efficacité des flux - rentabilité	- Agrégation de métriques de performance - indicateurs issus des SI métiers - méthodes multicritères	- Aide à la décision opérationnelle et stratégique - arbitrage coût–performance
Sociale (Safitri *et al.,* 2018), (Docekalova, 2017)	- Conditions de travail - sécurité - équité des processus - aspects organisationnels	- Indicateurs qualitatifs - enquêtes - logique floue - agrégation experte	- Évaluation globale de la durabilité organisationnelle - diagnostic non automatisé
Technologique (QBL) (Watanabe *et al.,* 2017)	- Niveau d'automatisation - intégration des SI, capacités de monitoring en temps réel	- Évaluation multicritères - intégration dans des cadres hiérarchiques (ex. ISA-95)	- Pilotage des systèmes productifs durables - soutien à la transformation numérique

Ce tableau souligne un déséquilibre en faveur des dimensions environnementales et économiques, ainsi qu'une faible opérationnalisation des indicateurs sociaux dans les systèmes d'information. Les indicateurs environnementaux, tels que la consommation d'énergie, les émissions de gaz à effet de serre ou la gestion des déchets, sont les plus fréquemment mobilisés, en particulier dans les contextes industriels (Watanabe *et al.*, 2017; Rai et Daniels, 2016; Kroeger *et al*., 2024). Les indicateurs économiques sont également présents, souvent sous la forme de coûts, de productivité ou de rentabilité, tandis que les indicateurs sociaux restent plus marginaux et difficilement opérationnalisables. Certains travaux proposent des méthodes d'agrégation et de pondération des indicateurs, notamment à l'aide de méthodes multicritères ou de normalisation par rapport à des objectifs cibles (Kroeger

et al., 2024; Watanabe *et al.,* 2017). Ces approches permettent de produire des scores synthétiques facilitant la comparaison de scénarios ou d'alternatives décisionnelles. Toutefois, elles reposent sur des choix de pondération qui reflètent des priorités organisationnelles et peuvent influencer fortement les résultats. Une contribution notable est l'introduction d'une dimension technologique supplémentaire, conduisant à une approche de type Quadruple Bilan (Quadruple Bottom Line) qui considère les capacités numériques et la maturité des SI comme un levier de durabilité (Watanabe *et al.,* 2017). Cette perspective reste toutefois peu développée dans la littérature et ouvre des pistes intéressantes pour la recherche en SI. Plusieurs travaux mobilisent des indicateurs agrégés sans exploiter pleinement la structure des processus sous-jacents. Cela atténue l'intérêt de la fouille de processus, en rapprochant certaines approches de méthodes analytiques plus classiques, centrées sur des agrégations de données plutôt que sur l'analyse des dynamiques d'exécution.

5.4. RQ4 – Enjeux sociaux et éthiques

L'analyse met en évidence un décalage important entre l'importance théorique des enjeux sociaux et éthiques associés à l'application de la fouille de processus à la durabilité et leur prise en compte effective dans les travaux étudiés.

Les questions d'équité, de confidentialité et de transparence sont principalement abordées dans des travaux conceptuels ou méthodologiques sur la Green Data Science et la fouille de processus responsable (van der Aalst, 2016). En revanche, elles sont rarement intégrées de manière explicite dans les cadres analytiques appliqués à la durabilité. La majorité des études se concentre sur l'amélioration des performances environnementales ou économiques, sans analyser les effets potentiels des décisions issues de la fouille sur les individus ou les groupes concernés. L'actionnabilité des résultats constitue un autre enjeu clé. Si certains travaux proposent des mécanismes explicites reliant les analyses de fouille de processus à des décisions opérationnelles ou stratégiques (Kroeger *et al.,* 2024; Kumbhar *et al.*, 2022; Park et van der Aalst, 2022), d'autres se limitent à des analyses descriptives sans lien clair avec les processus décisionnels. Cela limite l'impact réel des approches sur la durabilité des SI.

Le Tableau 3 synthétise les principaux enjeux sociaux et éthiques associés à l'application de la fouille à la durabilité et leur prise en compte dans les travaux analysés. Il met en évidence un décalage entre l'identification conceptuelle de ces enjeux et leur opérationnalisation dans les SI décisionnels. L'analyse révèle que les dimensions sociales et éthiques restent largement sous-explorées dans les applications de la fouille à la durabilité. Leur intégration systématique apparaît comme un axe de recherche prioritaire pour concevoir des SI décisionnels capables de soutenir une durabilité à la fois opérationnelle, stratégique et responsable.

Tableau 3 – Enjeux sociaux et éthiques

Enjeu social / éthique	**Enjeux**	**Manifestation dans la fouille de processus**	**Prise en compte dans les travaux**
Equité	Risque de biais ou de discrimination	- découverte de processus biaisée	- largement absente des cadres appliqués

(van der Aalst, 2016)	induits par l'analyse des données de processus	- décisions automatisées défavorables à certains groupes	- traitée plutôt au niveau conceptuel
Confidentialité (van der Aalst, 2016)	Protection des données sensibles issues des SI et des logs de processus	- réidentification possible via horodatages - identifiants ou corrélations	- peu explicitée dans les études de durabilité - supposée mais rarement opérationnalisée
Transparence (Watanabe *et al.,* 2017), (Rai et Daniels, 2016), (Kroeger *et al.,* 2024), (Gribaudo et Manini, 2020)	Compréhensibilité et traçabilité des analyses et décisions issues de la fouille de processus	- modèles interprétables - diagnostics explicites - traçabilité des décisions	Relativement bien prise en compte dans les cadres industriels et de simulation
Actionnabilité (Kroeger *et al.,* 2024), (Kumbhar *et al.,* 2022), (Park et van der Aalst, 2022)	Capacité à transformer les résultats analytiques en décisions concrètes	- lien explicite entre indicateurs - contraintes et actions	- variable selon les travaux - formalisée dans les approches orientées décision

6. Discussion

L'analyse des travaux met en évidence le potentiel de la fouille de processus comme levier structurant pour l'aide à la décision en matière de durabilité, mais elle révèle surtout des différences fortes dans les choix de conception des systèmes analysés. Au-delà du constat désormais classique du déséquilibre en faveur des dimensions environnementales, les résultats montrent que ce déséquilibre est étroitement lié aux types de données mobilisées, aux cadres de modélisation retenus et aux formes d'aide à la décision effectivement proposées. Cette lecture permet de dépasser une analyse thématique pour mettre en évidence des mécanismes structurels propres aux SI décisionnels.

Nous pouvons noter la forte hétérogénéité des approches, tant sur les sources de données que sur les cadres de modélisation et les systèmes d'indicateurs mobilisés. Cela reflète la pluralité des contextes d'application, mais complique l'identification de cadres analytiques génériques et réutilisables pour la conception de SI décisionnels durables. Le cloisonnement entre approches organisationnelles et approches industrielles limite souvent la capacité à relier les décisions stratégiques aux impacts opérationnels réels. Les travaux proposant des cadres intégrés apparaissent prometteurs, mais restent encore minoritaires et coûteux à mettre en œuvre. La Figure 3 propose une synthèse des principaux apports et limites des approches analysées.

Un second point de discussion concerne la centralité des indicateurs environnementaux dans les applications de la fouille à la durabilité. Les indicateurs sociaux sont rarement intégrés de manière opérationnelle, en raison de leur caractère qualitatif, contextuel et difficilement mesurable à partir de données de processus.

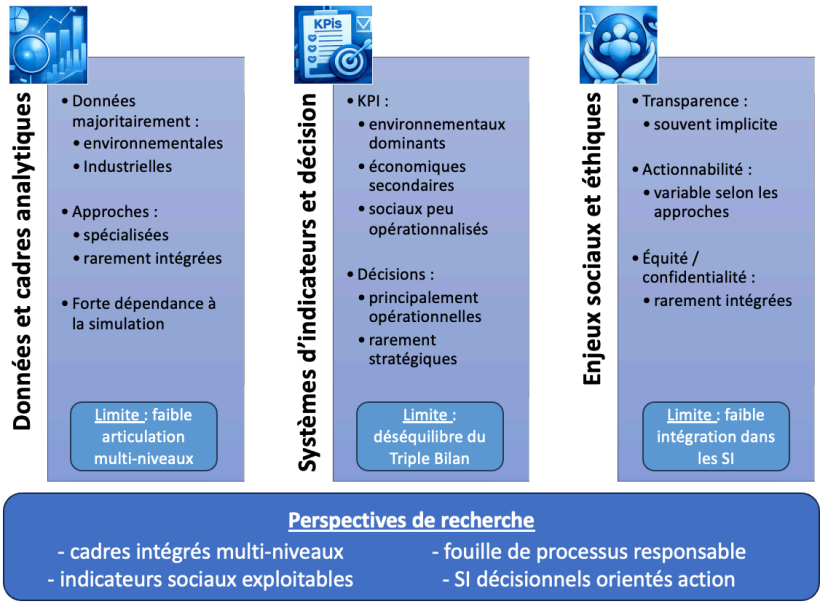


Figure 3 – Synthèse des apports et limites des approches de fouille de processus pour la durabilité.

Cette situation révèle une limite structurelle des approches actuelles, qui tendent à privilégier les dimensions de la durabilité les plus facilement quantifiables.

Par ailleurs, la discussion met en lumière le rôle encore marginal accordé aux enjeux éthiques et sociaux de la fouille de processus, notamment en termes d'équité, de confidentialité et de transparence. Bien que ces questions soient identifiées dans la littérature méthodologique, elles restent peu intégrées dans les cadres appliqués à la durabilité. Cette lacune est problématique pour la recherche en systèmes d'information, dans la mesure où l'aide à la décision fondée sur des données de processus peut produire des effets différenciés sur les acteurs et renforcer des asymétries existantes. L'intégration explicite de ces dimensions constitue ainsi un enjeu majeur pour le développement de SI réellement durables.

Un point central mis en évidence par cette analyse concerne la sous-exploitation des capacités propres à la fouille de processus. Alors que celle-ci permet d'analyser les processus réels à partir de leurs traces, notamment en termes de variantes d'exécution, de séquences d'activités ou de dynamiques temporelles, plusieurs travaux se limitent à des analyses agrégées. Cette situation réduit la valeur ajoutée spécifique de la fouille de processus et interroge son positionnement par rapport à des approches analytiques plus traditionnelles.

Enfin, l'analyse souligne l'importance de l'actionnabilité des résultats de la fouille de processus. Les approches les plus pertinentes pour l'aide à la décision sont celles qui établissent un lien explicite entre les analyses produites et les choix organisationnels ou opérationnels. À l'inverse, les travaux se limitant à des analyses descriptives peinent à démontrer leur valeur pour le pilotage de la durabilité. La fouille ne doit pas être envisagée uniquement comme un outil analytique, mais comme un composant des SI décisionnels.

Ce travail montre un décalage entre le potentiel théorique de la fouille de processus pour soutenir la durabilité et la maturité des cadres effectivement proposés. Ce décalage ouvre plusieurs perspectives de recherche, autour de la conception de cadres intégrés, de l'opérationnalisation des dimensions sociales et de l'intégration de principes éthiques dès la phase de conception des SI. Ces résultats soulignent également le rôle central de l'architecture des systèmes d'information dans la capacité à intégrer, transformer et exploiter les données de processus pour la durabilité.

7. Conclusion

Cet article propose une revue analytique structurée des travaux mobilisant la fouille de processus au service de la durabilité et de l'aide à la décision dans les SI. Nous proposons une grille d'analyse opérationnelle permettant de comparer les approches existantes et de guider la conception de systèmes d'information décisionnels orientés durabilité. Cela permet de structurer la littérature autour de quatre axes principaux : la constitution des données, les cadres de modélisation et d'évaluation, les systèmes d'indicateurs et les enjeux sociaux et éthiques. Les résultats montrent que la fouille de processus constitue un levier pertinent pour relier l'exécution réelle des processus aux objectifs de durabilité, en particulier dans les contextes industriels et organisationnels complexes. Toutefois, la littérature actuelle demeure fragmentée, avec une prédominance des approches environnementales et une prise en compte encore limitée des dimensions sociales, éthiques et technologiques de la durabilité. Cette situation limite la capacité des SI à soutenir une prise de décision véritablement intégrée et responsable.

Ce travail invite à repenser la fouille de processus non seulement comme un outil d'analyse des processus, mais comme un composant structurant de SI décisionnels durables. Il met en évidence la nécessité d'intégrer dès la conception des SI les choix relatifs aux données, aux indicateurs et aux principes éthiques, afin de garantir une aide à la décision à la fois actionnable et responsable. Ces résultats fournissent des points d'appui concrets pour la recherche en SI souhaitant concevoir des cadres analytiques réutilisables et alignés avec les enjeux contemporains de la durabilité.

Bibliographie

Acerbi F., P. W., Q. (2022). Fostering Circular Manufacturing through the integration of Genetic Algorithm and Process Mining. *Advances in Production Management Systems. Smart Manufacturing and Logistics Systems: Turning Ideas into Action*.

van der Aalst W. M. (2016). Green Data Science – Using Big Data in an "Environmentally Friendly" Manner. *International Conference on Enterprise Information Systems*.

Bauer J. (2024). *How process mining can help find a sustainability sweet spot*. Interview CELONIS, janvier 2024.

Docekalová M. P., D. M., A. K. (2017). Evaluations of corporate sustainability indicators based on fuzzy similarity graphs. *Ecological Indicators*, Elsevier.

Gribaudo M. , Manini D. (2020). Circular Economy: A Performance Evaluation Perspective. *ACM International Conference Proceedings*.

Graves N., K., W., van der Aalst W. M. (2023). ReThink Your Processes! A Review of Process Mining for Sustainability. *9th International Conference on ICT for Sustainability (ICT4S 2023)*, Rennes.

Joas A., Gierlich.-Joas M.. Bahr C., Bauer J. (2024). Towards Leveraging Process Mining for Sustainability – An Analysis of Challenges and Potential Solutions. *Business Process Management Forum*.

Kumbhar M., Ng A. HC., Bandaru S. (2022). Bottleneck Detection Through Data Integration, Process Mining and Factory Physics-Based Analytics. *Advances in Transdisciplinary Engineering*.

Kroeger S., Streibel L., Jordan P. Klages B. (2024). Sustainability assessment of production networks using simulation-data-based process mining. *Procedia Computer Science*.

Rai S., Daniels M. (2016). An event-log analysis and simulation-based approach for quantifying sustainability metrics in production facilities. *Winter Simulation Conference*.

Safitri, L. N., Sarno R., Budiawati G.I. (2018). *Improving Business Process by Evaluating Enterprise Sustainability Indicators Using Fuzzy Rule-Based Classification*. *International Seminar on Application for Technology of Information and Communication.*

van Dongen B. F., van der Aalst W. M., L. W. (2009). Process Mining: Overview and Outlook of Petri Net Discovery Algorithms. *Transactions on Petri Nets and Other Models*

Park G., van der Aalst W. M. (2022). Action-oriented process mining: bridging the gap between insights and actions. *Progress in Artificial Intelligence*.

Rabbanee, F. K., Hassan K., Taufique K.M.R., Shafiullah G.M., Dewan A. (2023). *Sustainability of Sustainable Business Practices: Challenges and Innovations*. *Sustainability*.

United Nations (2023). *Global Sustainable Development Report (GSDR)*.

United Nations (1992). *United Nations Conference on Environment and Development, Rio de Janeiro, Brazil, 3–14 June 1992*.

Union européenne (2024). *Directive (UE) 2022/2464 du Parlement européen et du Conseil du 14 décembre 2022 relative au reporting de durabilité des entreprises*.

Watanabe E. Marinho Da Silva R., Tsuzuki M., Junqueira F., Santos Filho D.J., Miygi P.E. (2017). Assessment of Sustainability for Production Control Based on Petri Net and Cyber-Physical Cloud System. *IFAC PapersOnLine*, Elsevier.

Fouille de personas via fouille de processus et apprentissage automatique non-supervisé

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. (2025). Persona Mining Using Process Mining and Unsupervised Machine-Learning, KES 2025, Osaka Japan.

1. Introduction

Dans un contexte de transformation numérique généralisée, les systèmes d'information et plateformes numériques sont omniprésents et soutiennent des domaines variés tels que l'éducation, la santé, le commerce électronique ou encore les écosystèmes organisationnels. Malgré cette omniprésence, l'expérience utilisateur (User eXperience – UX) demeure souvent insatisfaisante en raison de conceptions peu intuitives, de données incomplètes ou redondantes, et de difficultés d'accès à l'information pertinente. Ces limitations peuvent affecter négativement l'engagement, la satisfaction et la fidélité des utilisateurs (Sauro, Lewis, 2016).

L'amélioration de l'UX repose en grande partie sur la capacité des systèmes à proposer des recommandations personnalisées, adaptées aux besoins, aux intentions et au contexte des utilisateurs (Fernández-Portillo *et al*., 2024). Dans cette perspective, la notion de persona joue un rôle central. Un persona représente un type d'utilisateur caractérisé par des comportements, des objectifs et des attentes similaires. Traditionnellement, les personas sont construits à partir d'enquêtes, d'entretiens ou de questionnaires, méthodes qui sont coûteuses, intrusives et parfois biaisées.

Nous partons du constat que les traces d'utilisation laissées par les utilisateurs dans les systèmes, sous forme de journaux d'événement (event logs), constituent une source de données riche, objective et facilement accessible pour analyser les comportements réels. Cependant, la littérature existante en fouille de comportement (behavior mining) ne propose pas de méthodes permettant de construire des personas directement à partir de ces event logs. Ainsi, l'objectif principal de l'article est de proposer une approche innovante pour identifier et construire automatiquement des personas à partir des traces d'activité des utilisateurs, en combinant le process mining et des techniques d'apprentissage automatique non supervisé.

2. Fouille de personas basée sur l'apprentissage non-supervisé

L'approche proposée vise à extraire des personas à partir de journaux d'événements générés par les interactions des utilisateurs avec un système. Les event logs

doivent contenir au minimum trois attributs essentiels: 1) un identifiant de cas (USE CASE ID), 2) une activité (ACTIVITY) et 3) un horodatage (TIMESTAMP). Quelle que soit la forme initiale des données (CSV, XES, TXT), celles-ci sont prétraitées et normalisées sous forme de DataFrame afin de faciliter les étapes suivantes.

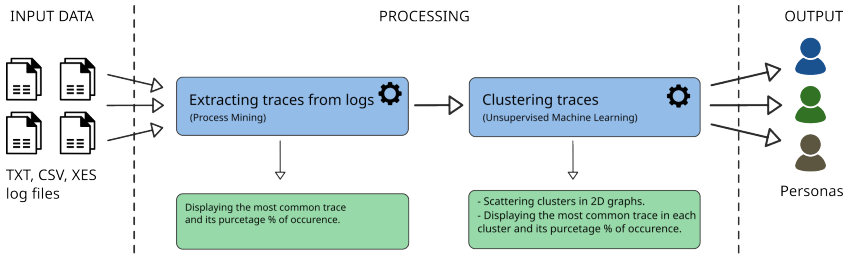


FIGURE 1 – Vue globale de la méthode proposée

Comme le montre la figure 1, à partir des event logs, nous appliquons des techniques de process mining pour extraire des traces, c'est-à-dire des séquences chronologiques d'activités réalisées par un utilisateur au cours d'un cas donné. Chaque trace représente un comportement utilisateur complet, depuis le début jusqu'à la fin d'une interaction ou d'une session. Ces traces peuvent être représentées sous forme de graphes dirigés, reflétant l'ordre temporel des événements. Nous observons une grande diversité de traces, allant de séquences simples à des enchaînements complexes, avec des répétitions d'activités possibles. Une première analyse consiste à identifier les traces les plus fréquentes et leur taux d'occurrence dans le jeu de données.

La seconde étape consiste à regrouper les traces similaires afin d'identifier des comportements proches. Pour cela, nous utilisons le clustering non supervisé, en particulier l'algorithme k-means, appliqué aux représentations des traces. Chaque cluster correspond alors à un ensemble de traces similaires, interprété comme une classe de comportement ou un persona.

Afin de déterminer le nombre optimal de clusters, et donc de personas, deux métriques de validation sont utilisées: le Silhouette Score (Shahapure, Nicholas, 2020) et le Davies-Bouldin Index (Xiao *et al.*, 2017). Ces métriques permettent d'évaluer respectivement la cohésion interne des clusters et leur séparation. L'analyse de ces indicateurs guide le choix du nombre de personas pertinents pour chaque jeu de données. Une fois les clusters établis, chaque utilisateur est associé à un persona, ce qui permet de relier les comportements observés à des profils d'utilisateurs exploitables dans des systèmes de recommandation.

3. Évaluation expérimentale

L'approche est évaluée sur trois jeux de données, dont deux sont décrits en détail dans l'article : 1) Permit Log Dataset: issu du BPI Challenge 2020 (ICPM), ce jeu de données concerne des demandes d'autorisation de déplacement dans une organisation.

Il contient plus de 86 000 événements et présente une forte variabilité des comportements utilisateurs. 2) Call Center Dataset : provenant des logs de démonstration de Fluxicon Disco, ce dataset décrit les interactions entre clients et un centre d'appel, incluant redirections, rappels et clôtures d'appels. Ces deux jeux de données respectent les prérequis définis pour l'extraction des traces et permettent de tester la robustesse de l'approche dans des contextes différents.

Les résultats montrent qu'il est possible d'identifier des clusters stables et interprétables à partir des traces extraites. Pour le dataset Permit Log, les métriques Silhouette et Davies-Bouldin indiquent que trois clusters offrent le meilleur compromis entre cohésion et séparation. Pour le dataset Call Center, les résultats sont plus nuancés, mais suggèrent que quatre clusters constituent un choix pertinent.

Les visualisations en deux dimensions confirment l'existence de groupes bien distincts, chacun correspondant à un type de comportement utilisateur spécifique. L'analyse qualitative des traces dominantes dans chaque cluster révèle des différences significatives dans les parcours utilisateurs, confirmant que chaque cluster peut être interprété comme un persona distinct.

4. Conclusion

L'étude propose une approche innovante et non intrusive pour le mining de personas basée sur l'analyse de traces d'utilisation réelles. En combinant process mining et apprentissage non supervisé, nous démontrons qu'il est possible de construire automatiquement des personas fiables à partir d'event logs, sans recourir à des enquêtes ou questionnaires utilisateurs. Cette contribution ouvre des perspectives importantes pour les systèmes de recommandation et l'orchestration d'entités dans les écosystèmes numériques, en intégrant le comportement réel des utilisateurs dans les processus de personnalisation.

Les travaux futurs envisagent d'enrichir cette approche par des données qualitatives (enquêtes, interviews), d'expérimenter d'autres algorithmes de clustering (DBSCAN, clustering hiérarchique, modèles de mélanges gaussiens) et d'évaluer l'impact concret des personas sur la performance des recommandations dans des écosystèmes numériques réels.

Bibliographie

Fernández-Portillo A., Ramos-Vecino N., Ramos-Mariño A., Cachón-Rodríguez G. (2024). How the digital business ecosystem affects stakeholder satisfaction: its impact on business performance. *Review of Managerial Science*, p. 1–20.

Sauro J., Lewis J. R. (2016). *Quantifying the user experience: Practical statistics for user research*. Morgan Kaufmann.

Shahapure K. R., Nicholas C. (2020). Cluster quality analysis using silhouette score. , vol. 1, n° 1, p. 747-748.

Xiao J., Lu J., Li X. (2017). Davies bouldin index based hierarchical initialization k-means. *Intelligent Data Analysis*, vol. 21, n° 6, p. 1327–1338.

Piloter la Durabilité par la Fouille de Processus

Cadres d'Analyse et Indicateurs pour l'Aide à la Décision dans les Systèmes d'Information

Nikita Valenza, Rébecca Deneckère

Centre de Recherche en Informatique
Université Paris 1 Panthéon-Sorbonne, France
nikita.valenza@yahoo.com, rebecca.deneckere@univ-paris1.fr

RESUME. La durabilité est un enjeu central pour les systèmes d'information décisionnels, en particulier dans des contextes organisationnels et industriels soumis à de fortes contraintes environnementales, économiques et sociales. La fouille de processus s'impose comme une approche prometteuse pour relier l'exécution réelle des processus aux objectifs de durabilité, à partir de l'analyse de traces issues des systèmes d'information. Cet article propose une revue analytique structurée des travaux mobilisant la fouille de processus pour l'évaluation et le pilotage de la durabilité. Les résultats mettent en évidence une forte hétérogénéité des approches, tant sur les sources de données que sur les cadres conceptuels, ainsi qu'un déséquilibre marqué entre les dimensions environnementales et sociales. L'article discute les implications de ces constats pour la conception de systèmes d'information décisionnels durables et identifie plusieurs pistes de recherche, notamment autour de l'intégration d'indicateurs éthiques et de l'actionnabilité des résultats.

ABSTRACT. Sustainability is a key concern for decision-support information systems, particularly in organizational and industrial contexts facing strong environmental, economic, and social constraints. Process Mining emerges as a promising approach to connect actual process execution with sustainability objectives through the analysis of event logs extracted from information systems. This paper presents a structured analytical review of research works applying Process Mining to sustainability assessment and decision support. The results highlight a strong diversity of approaches in terms of data sources and conceptual frameworks, as well as a marked imbalance in favor of environmental dimensions compared to social aspects. The paper discusses the implications of these findings for the design of sustainable decision-support information systems and outlines several research perspectives, particularly related to ethical indicators and the actionability of Process Mining outcomes.

MOTS-CLES : Fouille de processus, Durabilité, Systèmes d'Information Décisionnels, Indicateurs de Performance, Aide à la Décision.

KEYWORDS: Process Mining, Sustainability, Decision-support Information Systems, Performance Indicators, Decision-making

1. Introduction

La durabilité constitue un enjeu structurant pour les organisations, impliquant l'intégration conjointe de dimensions environnementales, économiques et sociales dans les systèmes d'information décisionnels (Rabbanee *et al.,* 2023), (Union Européenne, 2024). Dans ce contexte, la fouille de processus permet d'exploiter les traces issues des systèmes d'information afin d'analyser l'exécution réelle des processus et de soutenir la prise de décision.

Plusieurs travaux récents mettent en évidence le potentiel de la fouille de processus pour analyser des dimensions environnementales de la durabilité, notamment la consommation d'énergie, les émissions de gaz à effet de serre ou l'optimisation des flux de production (Watanabe *et al.,* 2017; Rai et Daniels, 2016; Kroeger *et al.,* 2024). D'autres approches étendent cette perspective à des cadres d'évaluation plus larges, intégrant des dimensions économiques et sociales afin d'évaluer la durabilité globale de systèmes productifs ou de processus métiers dans une logique de Triple Bilan (Triple Bottom Line) (Kroeger *et al.,* 2024; Docekalova, 2017). Toutefois, malgré cet intérêt croissant, la littérature demeure hétérogène et fragmentée. Les travaux diffèrent fortement selon les types de données mobilisées, les cadres de modélisation retenus, les méthodes de construction des indicateurs et les niveaux d'analyse considérés, qu'ils soient organisationnels, industriels ou intégrés. Cette diversité rend difficile l'identification de cadres analytiques cohérents permettant de guider la conception de SI décisionnels durables. En outre, la majorité des travaux se concentre sur les dimensions environnementales et économiques, tandis que les enjeux sociaux et éthiques liés à l'exploitation des données de processus, tels que l'équité[1], la confidentialité ou la transparence des analyses, restent encore marginalement abordés (van der Aalst, 2016; Park et van der Aalst, 2022).

Ces constats soulèvent une question centrale pour la recherche en systèmes d'information : comment la fouille de processus peut-elle être mobilisée de manière structurée pour soutenir l'aide à la décision en matière de durabilité au sein des SI ? Cette question dépasse la simple application de techniques analytiques à des indicateurs environnementaux. Elle interroge la manière dont les données de processus sont constituées, modélisées et interprétées, ainsi que les cadres conceptuels et décisionnels dans lesquels elles s'inscrivent, afin de produire des résultats exploitables, responsables et alignés avec les objectifs de durabilité. Pour répondre à cette question, cet article propose une revue analytique structurée des travaux mobilisant la fouille de processus au service de la durabilité et de l'aide à la décision. Contrairement aux revues existantes majoritairement orientées vers des cadres conceptuels ou sectoriels, ce travail adopte une grille de lecture explicitement ancrée dans les systèmes d'information décisionnels, en analysant conjointement la constitution des données, les cadres d'évaluation et les mécanismes de traduction des

[1] Dans la littérature anglo-saxonne en fouille de processus et en science des données, le terme *fairness* désigne des propriétés formalisables liées à l'absence de biais et de discrimination dans les données, les modèles ou les résultats algorithmiques ; il est ici rapproché du terme français *équité*, entendu au sens opérationnel et non uniquement normatif.

résultats analytiques en décisions. L'objectif est d'identifier les choix structurants et les angles morts de la littérature afin de guider la conception de SI décisionnels durables.

L'article est structuré comme suit. La section 2 décrit quelques fondamentaux et la section 3 les travaux connexes. La méthodologie de recherche est expliquée en section 4. La section 5 propose les résultats de l'analyse organisée selon plusieurs axes. Nous proposons une discussion en section 6 et nous concluons en section 7.

2. Fondamentaux

Cette section présente les fondements théoriques mobilisés dans cet article et introduit le cadre qui structure l'analyse des travaux étudiés. Elle vise à clarifier les notions de fouille de processus, de durabilité et de SI décisionnels, ainsi que leurs articulations, afin de fournir une base cohérente pour l'analyse comparative.

Fouille de processus et systèmes d'information. La fouille de processus désigne un ensemble de techniques visant à analyser l'exécution réelle des processus à partir des traces enregistrées. Ces traces associent des informations telles qu'un identifiant de cas, une activité, un horodatage et, selon les contextes, des attributs complémentaires liés aux ressources ou aux objets manipulés (Bauer, 2024; van Dongen et van der Aalst, 2009). À la différence des approches traditionnelles de modélisation des processus, la fouille ne repose pas sur des modèles prescriptifs, mais sur l'observation empirique des comportements effectifs des systèmes. Les ERP, MES ou systèmes de production automatisés génèrent des volumes importants de données événementielles, qui peuvent être exploitées pour reconstruire des modèles de processus, analyser des performances ou vérifier la conformité des pratiques (Bauer, 2024). La fouille de processus est donc un moyen de transformer des données opérationnelles en connaissances exploitables pour la prise de décision. Les principaux types de techniques de fouille de processus incluent la découverte de processus, l'analyse de conformité et l'analyse de performance (van Dongen et van der Aalst, 2009). La découverte de processus vise à produire des modèles représentant les enchaînements d'activités observés, souvent sous forme de réseaux de Petri ou de variantes dérivées. L'analyse de conformité permet de comparer ces modèles aux processus prescrits ou attendus, tandis que l'analyse de performance exploite les dimensions temporelles et quantitatives des traces afin d'identifier des inefficiences ou des goulots d'étranglement. Ces techniques constituent le socle analytique sur lequel reposent les applications de la fouille de processus à des objectifs de durabilité.

Durabilité et SI décisionnels. La durabilité est généralement appréhendée à travers la notion de Triple Bilan (United Nations, 1992; Rabbanee *et al.,* 2023). Pour les systèmes d'information, cela se traduit par la nécessité de concevoir des outils capables de mesurer, d'analyser et de piloter ces différentes dimensions de manière intégrée. Les SI décisionnels jouent ici un rôle clé, en agrégeant des données hétérogènes et en produisant des indicateurs destinés à soutenir les choix stratégiques et opérationnels des organisations. Les indicateurs de durabilité, souvent formalisés sous forme de KPI, constituent un élément central de ces systèmes. Ils permettent de traduire des objectifs abstraits, tels que la réduction de l'empreinte carbone ou

l'amélioration des conditions de travail, en mesures opérationnelles exploitables (Docekalova, 2017). Toutefois, la littérature souligne la difficulté de construire des systèmes d'indicateurs cohérents, comparables et adaptés aux contextes organisationnels, en raison de la diversité des sources de données, des niveaux d'analyse et des référentiels mobilisés (Docekalova, 2017; United Nations, 2023). Dans ce contexte, les SI décisionnels orientés durabilité doivent répondre à plusieurs exigences. Ils doivent être capables d'intégrer des données issues de processus opérationnels, de supporter des analyses multicritères et de produire des résultats interprétables par les décideurs. Ils doivent également prendre en compte des contraintes éthiques et réglementaires, notamment en matière de confidentialité et de transparence, afin de garantir la légitimité des décisions produites (van der Aalst, 2016).

Articulation entre fouille de processus et durabilité. L'application de la fouille de processus à la durabilité repose sur l'hypothèse selon laquelle les impacts environnementaux, économiques et sociaux des organisations sont en grande partie déterminés par l'exécution concrète de leurs processus. En analysant les processus tels qu'ils sont réellement réalisés, la fouille de processus offre la possibilité de relier des indicateurs de durabilité à des activités, des enchaînements ou des configurations organisationnelles spécifiques (Graves et van der Aalst, 2023; Kroeger *et al.,* 2024). La littérature montre que cette articulation peut prendre des formes variées. Dans certains cas, la fouille est utilisée pour enrichir des analyses existantes en y intégrant des indicateurs environnementaux ou économiques (Watanabe *et al.,* 2017; Rai et Daniels, 2016). Dans d'autres cas, il constitue le cœur d'un cadre analytique plus large, combinant modélisation des processus, simulation et construction de systèmes d'indicateurs orientés durabilité (Kroeger *et al.,* 2024; Gribaudo et Manini, 2020). Cependant, la construction de traces adaptées à l'analyse de la durabilité nécessite souvent de combiner des données hétérogènes, issues à la fois de SI métier et de sources externes, telles que des bases de données environnementales ou des référentiels normatifs. De plus, le choix des indicateurs et des cadres d'évaluation influence fortement les résultats produits et leur interprétation par les décideurs, ce qui renforce la nécessité d'une approche structurée et explicite.

Cadre de travail de l'analyse. Ce travail d'analyse adopte un cadre articulant trois niveaux complémentaires, correspondant aux principales décisions de conception d'un SI décisionnel orienté durabilité. Le premier niveau porte sur les données et la constitution des traces, conditionnant la mesurabilité des impacts. Dans ce travail, les systèmes d'information industriels (MES, CPS, systèmes de capteurs) sont considérés comme des SI à part entière, opérant à un niveau opérationnel et cyber-physique, et complémentaires des systèmes d'information décisionnels de l'entreprise. Le deuxième concerne les cadres de modélisation et d'évaluation, qui structurent l'interprétation des processus. Le troisième porte sur les systèmes d'indicateurs et l'aide à la décision, où se matérialise la valeur décisionnelle des analyses. Ce cadre permet ainsi d'analyser les travaux non seulement selon leurs résultats, mais selon les choix de conception qu'ils impliquent pour les SI. La Figure 1 illustre le rôle de la fouille comme médiateur entre les SI et l'aide à la décision orientée durabilité, en mettant en évidence les niveaux de données, d'analyse et de décision, ainsi que le caractère transversal des enjeux sociaux et éthiques.

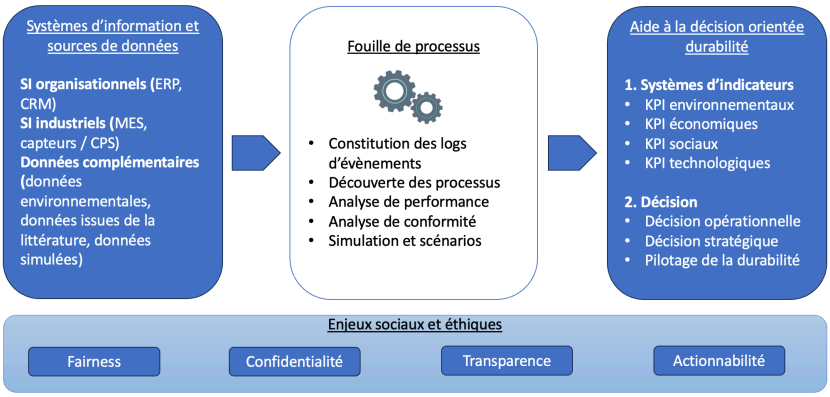


Figure 1 – Cadre de la fouille de processus au service de la durabilité dans les SI.

Cette figure met en évidence l'articulation entre les sources de données, les analyses de processus et la construction de systèmes d'indicateurs, ainsi que le caractère transversal des enjeux sociaux et éthiques.

Afin d'illustrer concrètement l'articulation entre données de processus, indicateurs de durabilité et aide à la décision, considérons un processus de production industrielle. Des données issues de capteurs peuvent être associées aux activités du processus, par exemple en termes de consommation énergétique. Après constitution des traces et enrichissement sémantique, la fouille de processus permet d'identifier différentes variantes d'exécution. L'analyse peut alors révéler que certaines variantes sont plus énergivores que d'autres, en raison de séquences d'activités ou de temps d'attente spécifiques. Cette capacité à relier directement des indicateurs de durabilité à des configurations réelles de processus illustre le rôle de la fouille comme médiateur entre données opérationnelles et décision, en identifiant des leviers d'optimisation ciblés.

3. Travaux Connexes

Les travaux mobilisant la fouille de processus dans une perspective de durabilité s'inscrivent à l'intersection de plusieurs champs de recherche, notamment le Business Process Management, les SI décisionnels et les études sur la durabilité.

Application de la fouille de processus pour l'analyse et l'optimisation des performances opérationnelles. Ces travaux ont structuré les techniques de découverte, de conformité et de performance à partir de logs d'événements (Bauer, 2024; van Dongen et van der Aalst, 2009). Bien que principalement orientées vers l'efficacité, la réduction des coûts ou la conformité réglementaire, ces approches constituent le socle méthodologique sur lequel se sont appuyées les applications ultérieures de la fouille à des problématiques environnementales et sociétales.

Intégration des préoccupations environnementales dans l'analyse des processus. Ces travaux mobilisent la fouille de processus pour analyser la consommation

énergétique, les émissions de gaz à effet de serre ou l'utilisation des ressources dans des contextes industriels et de production (Watanabe *et al.,* 2017; Rai et Daniels, 2016). Toutefois, ces approches restent souvent focalisées sur une dimension spécifique de la durabilité et ne proposent pas de cadre décisionnel intégrant l'ensemble des dimensions possibles.

Vision intégrée de la durabilité. Ces travaux mobilisent des cadres tels que le Triple Bilan qui renvoie à l'évaluation conjointe des différentes dimensions. Certaines contributions proposent des systèmes d'indicateurs combinant fouille de processus, simulation et méthodes d'agrégation afin d'évaluer la durabilité de systèmes productifs ou de réseaux de production (Kroeger *et al.,* 2024; Gribaudo et Manini, 2020). D'autres s'appuient sur des techniques telles que la logique floue pour traiter des indicateurs qualitatifs ou imprécis et comparer la performance durable de processus ou d'organisations (Docekalova, 2017; Safitri *et al.,* 2018). Ces approches élargissent le périmètre de l'analyse, mais restent hétérogènes dans leurs choix méthodologiques. Un travail de référence (Graves et van der Aalst, 2023) offre une vue d'ensemble des applications de la fouille à la durabilité et propose un cadre, le PM4S, visant à relier la fouille aux principes de l'économie circulaire. L'accent est mis sur la gestion des flux de ressources, la logistique inverse et l'usage de modèles orientés objets. Si ce travail fournit une base théorique solide et met en lumière le potentiel de la fouille pour soutenir des pratiques durables, il adopte principalement une perspective conceptuelle et sectorielle, centrée sur l'économie circulaire, sans proposer une analyse détaillée des méthodes de constitution des données, des cadres d'évaluation ou des systèmes d'indicateurs mobilisés dans les travaux étudiés.

Par ailleurs, les enjeux sociaux et éthiques associés à l'exploitation des données de processus sont encore peu présents dans les travaux existants, excepté (van der Aalst, 2016) qui aborde explicitement des questions d'équité, de confidentialité et de transparence dans l'analyse des données. De même, les travaux sur l'Action-Oriented Process Mining (Park et van der Aalst, 2022) mettent en évidence l'importance de l'actionnabilité des résultats et du passage des analyses à des décisions concrètes. Néanmoins, ces contributions restent rarement mobilisées de manière systématique dans les travaux appliquant la fouille de processus à la durabilité. La littérature met en évidence le potentiel de la fouille de processus pour soutenir des objectifs de durabilité, mais elle demeure fragmentée selon les dimensions de la durabilité considérées, les types de données exploitées et les cadres analytiques retenus. Peu de travaux proposent une grille de lecture transversale permettant d'analyser conjointement les méthodes de constitution des données, les cadres de modélisation, les systèmes d'indicateurs et les enjeux sociaux associés, dans une perspective explicitement orientée SI décisionnels. C'est précisément cette lacune que le présent travail se propose de combler, en offrant une revue structurée des approches et en mettant en évidence leurs implications pour la conception de SI durables.

4. Méthode de Recherche

Cet article adopte une démarche de revue analytique structurée visant à analyser de manière systématique les travaux mobilisant la fouille de processus pour soutenir

l'aide à la décision en matière de durabilité dans les systèmes d'information. La méthodologie repose sur une sélection de contributions scientifiques et sur une grille d'analyse construite à partir du cadre de travail présenté précédemment.

Définition de la question de recherche. La question de recherche principale, formulée dans l'introduction, est la suivante : **Comment la fouille de processus peut-elle être mobilisée de manière structurée pour soutenir l'aide à la décision en matière de durabilité au sein des systèmes d'information ?** Cette question est déclinée en quatre sous-questions de recherche, qui permettent d'examiner successivement les choix méthodologiques, conceptuels et décisionnels des approches étudiées : *RQ1 : Quelles sont les méthodes de constitution des données et des traces mobilisées pour la quantification d'indicateurs de durabilité à l'aide de la fouille de processus ? RQ2 : Quels cadres de modélisation et d'évaluation sont proposés pour intégrer la durabilité dans l'analyse des processus ? RQ3 : Comment les différentes perspectives de la durabilité influencent-elles la construction et l'usage des systèmes d'indicateurs pour l'aide à la décision ? RQ4 : Quels sont les enjeux sociaux et éthiques associés à l'application de la fouille de processus à la durabilité dans des environnements de données à grande échelle ?*

Sélection du corpus. La sélection du corpus s'appuie sur une recherche bibliographique structurée, réalisée sur la base de données Scopus. Les mots-clés ont été définis de manière à couvrir trois axes principaux : la fouille de processus, les indicateurs et métriques, ainsi que la durabilité. Ces axes ont été combinés à l'aide d'opérateurs logiques afin d'identifier des travaux traitant explicitement de l'articulation entre fouille et durabilité. La chaîne de recherche utilisée est la suivante: (TITLE-ABS-KEY("Process Mining") OR TITLE-ABS-KEY("Event Log") OR TITLE-ABS-KEY("Process Optimization") OR TITLE-ABS-KEY("Petri Nets")) AND (TITLE-ABS-KEY("Metrics") OR TITLE-ABS-KEY("Data Analytics") OR TITLE-ABS-KEY("Indicators") OR TITLE-ABS-KEY("Measurements") OR TITLE-ABS-KEY("Measure")) AND (TITLE-ABS-KEY("Sustainable Development Goals") OR TITLE-ABS-KEY("Sustainability") OR TITLE-ABS-KEY("Climate Action") OR TITLE-ABS-KEY("Corporate Social Responsibility") OR TITLE-ABS-KEY("Circular Economy")) AND PUBYEAR > 2015. Cette chaîne a permis d'identifier 161 sources possibles. La base Scopus a été retenue pour sa couverture large et sa structuration des métadonnées. L'absence d'intégration d'autres bases constitue une limite, mais ne remet pas en cause la diversité des approches identifiées.

Les critères d'inclusion sont les suivants : (i) articles scientifiques évalués par les pairs ; (ii) en anglais ou en français ; (iii) publiés après 2015, pour garantir la prise en compte des évolutions récentes de la fouille de processus et des enjeux de durabilité ; (iv) contributions mobilisant explicitement des techniques de fouille ou des approches assimilées à partir de traces ; (v) présence explicite d'indicateurs ou de métriques liés à la durabilité. Les critères d'exclusion comprennent les travaux portant sur l'extraction minière au sens géologique, les articles trop éloignés du champ des SI, ou ceux ne proposant pas d'analyse exploitable du point de vue des processus. Cette démarche a conduit à l'identification de 10 articles constituant le corpus analysé. Ce choix reflète un compromis entre exhaustivité et profondeur analytique, l'objectif de la revue n'étant pas de couvrir l'ensemble des publications liées à la durabilité, mais d'analyser en détail des travaux proposant une articulation explicite entre fouille de

processus, indicateurs et aide à la décision. La sélection privilégie la diversité des cadres analytiques et des niveaux d'analyse.

La démarche s'inspire des principes des revues de type PRISMA. Toutefois, elle ne vise pas l'exhaustivité statistique mais une analyse qualitative structurée. Certaines références non académiques sont mobilisées de manière complémentaire sans constituer le cœur du corpus analysé.

Démarche d'analyse. L'analyse repose sur une grille construite à partir du cadre présenté en section 2. Chaque contribution est analysée selon quatre axes correspondant aux sous-questions de recherche : les sources de données et les méthodes de constitution des traces, les cadres de modélisation et d'évaluation mobilisés, les systèmes d'indicateurs et leur rôle dans l'aide à la décision et les enjeux sociaux et éthiques abordés. Cela permet une comparaison systématique des approches étudiées, en mettant en évidence leurs points communs, leurs divergences et leurs limites. L'objectif n'est pas d'évaluer la performance relative des approches, mais d'identifier des tendances, des choix structurants et des lacunes potentielles. Il s'agit de mettre en évidence les choix de conception sous-jacents pour les SI décisionnels.

5. Analyse

Cette section présente l'analyse du corpus par sous-questions de recherche.

5.1. RQ1 – Méthodes de constitution des données et des traces

La première question de recherche porte sur les méthodes mobilisées pour constituer les données nécessaires à l'application de la fouille de processus dans une perspective de durabilité. L'analyse met en évidence une forte hétérogénéité des sources de données et des niveaux de granularité retenus. Le Tableau 1 propose une typologie des travaux analysés selon les types de données mobilisées et les niveaux d'analyse. Il positionne les travaux mobilisant explicitement des données de processus dans une perspective d'analyse (les contributions à dominante conceptuelle, qui ne se prêtent pas à une classification selon les types de données et les niveaux d'analyse, sont discutées séparément dans les travaux connexes et la discussion).

Une première distinction peut être établie entre les approches reposant principalement sur des données issues de SI organisationnels et celles mobilisant des données de niveau industriel. Les premières exploitent des traces extraites d'ERP ou de systèmes métiers afin d'analyser des processus à un niveau relativement agrégé, souvent orienté vers la gestion des flux, la conformité ou la performance organisationnelle (Docekalova, 2017; Park et van der Aalst, 2022). Ces approches facilitent l'intégration de considérations économiques et sociales, mais elles offrent une vision limitée des impacts environnementaux liés aux opérations physiques.

À l'inverse, les travaux centrés sur des environnements industriels exploitent des données de production à plus forte granularité, issues de MES, de capteurs ou de systèmes cyber-physiques (Watanabe *et al.,* 2017; Rai et Daniels, 2016; Kumbhar *et*

Tableau 1. Typologie des approches de fouille de processus

Type de données mobilisées / Niveaux	Organisationnel	Production / industriel	Intégré
Données organisationnelles de haut niveau logs ERP, données métier, enquêtes, entretiens	(Docekalova, 2017), (Safitri *et al.,* 2018), (Park et van der Aalst, 2022)		
Données opérationnelles de bas niveau MES, capteurs, logs machines, CPS		(Watanabe *et al.,* 2017), (Rai et Daniels, 2016), (Kumbhar *et al.,* 2022)	
Données hybrides / enrichies combinaison SI + données industrielles + littérature + simulation			(Kroeger *et al.,* 2024), (Gribaudo et Manini, 2020), (Acerbi, 2022)

al., 2022). Ces données permettent d'associer des indicateurs environnementaux, tels que la consommation énergétique ou les émissions, à des activités de processus spécifiques. Toutefois, leur intégration dans des traces exploitables pour la fouille de processus nécessite souvent des opérations complexes de prétraitement, de synchronisation et d'enrichissement des données.

Plusieurs travaux proposent des approches hybrides combinant données organisationnelles et industrielles, parfois enrichies par des données issues de la littérature ou de référentiels environnementaux (Kroeger *et al.,* 2024). Cette hybridation apparaît comme nécessaire pour construire des analyses de durabilité plus complètes, mais soulève des défis importants en termes de cohérence des données et de charge de mise en œuvre pour les SI.

Enfin, un nombre significatif de travaux recourt à la simulation, notamment à événements discrets, pour générer des données lorsque les traces réelles sont incomplètes ou indisponibles (Rai et Daniels, 2016; Kroeger *et al.,* 2024; Gribaudo et Manini, 2020). Ces approches permettent d'explorer des scénarios prospectifs.

Préparation des traces pour la fouille de processus. Au-delà des sources de données, l'exploitation de la fouille de processus repose sur des étapes de préparation des traces souvent peu explicitées dans les travaux analysés. Pourtant, ces transformations conditionnent directement la qualité des analyses produites. La construction de logs exploitables implique plusieurs opérations clés. Tout d'abord, un filtrage des données est nécessaire afin d'éliminer les événements incomplets, les doublons ou les traces bruitées. Ensuite, la structuration des cas constitue une étape centrale, consistant à définir ce qui constitue une instance de processus (par exemple une commande, un lot de production ou un cycle machine). Les données issues de sources hétérogènes doivent également être synchronisées temporellement et intégrées dans un format cohérent. Enfin, un enrichissement sémantique est souvent requis pour associer aux événements des attributs liés à la durabilité, tels que la consommation énergétique, les émissions ou l'utilisation de ressources. Ces étapes de transformation traduisent le passage de données brutes issues des SI vers des logs orientés processus, directement exploitables par les techniques de fouille. Leur

absence ou leur simplification limite fortement la capacité à analyser les processus réels et à produire des indicateurs pertinents. En particulier, la définition du cas et le choix du niveau de granularité influencent directement les modèles de processus découverts et, par conséquent, les indicateurs de durabilité qui peuvent être dérivés.

5.2. RQ2 – Cadres de modélisation et d'évaluation de la durabilité

Les cadres analysés peuvent être distingués selon leur capacité à exploiter explicitement la structure des processus, avec trois niveaux possibles : le niveau *activité*, où les indicateurs sont associés à des tâches individuelles, le niveau *trace*, où les analyses portent sur des instances complètes de processus, et le niveau des *variantes*, où les différentes trajectoires d'exécution sont comparées. Cette distinction est essentielle pour caractériser l'apport spécifique de la fouille de processus par rapport à des approches analytiques classiques.

L'analyse révèle également trois grandes familles d'approches.

Cadres orientés pilotage organisationnel. Ils se concentrent sur l'évaluation de processus métiers ou de chaînes de valeur à un niveau agrégé. Ils s'appuient sur des indicateurs issus des SI décisionnels et mobilisent des techniques telles que l'analyse multidimensionnelle ou la logique floue afin de traiter des données qualitatives ou imprécises (Docekalova, 2017; Safitri *et al.,* 2018). Ces approches sont particulièrement adaptées à des comparaisons inter-organisationnelles ou à des évaluations globales de la durabilité, mais elles offrent une visibilité limitée sur les mécanismes opérationnels sous-jacents.

Cadres orientés performance productive. Ils s'intéressent aux processus industriels et aux flux physiques. Ils mobilisent fréquemment des modèles de simulation, des réseaux de Petri ou des approches issues de la physique des systèmes de production pour analyser l'efficacité énergétique, les goulots d'étranglement ou l'utilisation des ressources (Watanabe *et al.,* 2017; Rai et Daniels, 2016; Kumbhar *et al.,* 2022). Ces cadres permettent une analyse fine des impacts environnementaux, mais ils sont souvent moins intégrés aux SI décisionnels de l'organisation.

Cadres systémiques ou intégrateurs. Ils combinent fouille, simulation et systèmes d'indicateurs afin d'évaluer la durabilité à plusieurs niveaux, du poste de travail au réseau de production (Kroeger *et al.,* 2024; Gribaudo et Manini, 2020; Acerbi, 2022). Ces approches apparaissent comme les plus prometteuses pour soutenir l'aide à la décision, mais elles impliquent une complexité technique et organisationnelle élevée, notamment en termes d'intégration des données et de gouvernance des modèles.

5.3. RQ3 – Systèmes d'indicateurs et aide à la décision

Les systèmes d'indicateurs identifiés peuvent être distingués selon leur ancrage dans les processus. Certains indicateurs sont calculés indépendamment de la structure des processus, tandis que d'autres sont explicitement liés aux activités, aux séquences

ou aux variantes d'exécution. Cette distinction permet d'évaluer dans quelle mesure les approches exploitent réellement les spécificités de la fouille de processus.

L'analyse montre que la majorité des travaux s'inscrit dans une logique de Triple Bilan, avec une prédominance marquée des indicateurs environnementaux. Le Tableau 2 synthétise les dimensions de la durabilité abordées dans les travaux analysés, les types d'indicateurs mobilisés et leur rôle dans l'aide à la décision.

Tableau 2 – Dimensions de la durabilité

Dimension de durabilité	**Types d'indicateurs**	**Méthodes de construction**	**Finalité décisionnelle**
Environnementale (Watanabe *et al.,* 2017), (Rai et Daniels, 2016), (Kroeger *et al.,* 2024), (Gribaudo et Manini, 2020), (Acerbi, 2022)	- Consommation énergétique - émissions de GES - flux de matières, déchets - efficacité énergétique	- Mesures directes à partir des logs - couplage fouille–simulation (DES) - normalisation par objectifs	- Optimisation des processus, réduction de l'empreinte environnementale - comparaison de scénarios
Économique (Kroeger *et al.,* 2024), (Kumbhar *et al.,* 2022), (Docekalova, 2017), (Park et van der Aalst, 2022)	- Coûts de production - productivité - taux d'utilisation - efficacité des flux - rentabilité	- Agrégation de métriques de performance - indicateurs issus des SI métiers - méthodes multicritères	- Aide à la décision opérationnelle et stratégique - arbitrage coût–performance
Sociale (Safitri *et al.,* 2018), (Docekalova, 2017)	- Conditions de travail - sécurité - équité des processus - aspects organisationnels	- Indicateurs qualitatifs - enquêtes - logique floue - agrégation experte	- Évaluation globale de la durabilité organisationnelle - diagnostic non automatisé
Technologique (QBL) (Watanabe *et al.,* 2017)	- Niveau d'automatisation - intégration des SI, capacités de monitoring en temps réel	- Évaluation multicritères - intégration dans des cadres hiérarchiques (ex. ISA-95)	- Pilotage des systèmes productifs durables - soutien à la transformation numérique

Ce tableau souligne un déséquilibre en faveur des dimensions environnementales et économiques, ainsi qu'une faible opérationnalisation des indicateurs sociaux dans les systèmes d'information. Les indicateurs environnementaux, tels que la consommation d'énergie, les émissions de gaz à effet de serre ou la gestion des déchets, sont les plus fréquemment mobilisés, en particulier dans les contextes industriels (Watanabe *et al.*, 2017; Rai et Daniels, 2016; Kroeger *et al*., 2024). Les indicateurs économiques sont également présents, souvent sous la forme de coûts, de productivité ou de rentabilité, tandis que les indicateurs sociaux restent plus marginaux et difficilement opérationnalisables. Certains travaux proposent des méthodes d'agrégation et de pondération des indicateurs, notamment à l'aide de méthodes multicritères ou de normalisation par rapport à des objectifs cibles (Kroeger

et al., 2024; Watanabe *et al.,* 2017). Ces approches permettent de produire des scores synthétiques facilitant la comparaison de scénarios ou d'alternatives décisionnelles. Toutefois, elles reposent sur des choix de pondération qui reflètent des priorités organisationnelles et peuvent influencer fortement les résultats. Une contribution notable est l'introduction d'une dimension technologique supplémentaire, conduisant à une approche de type Quadruple Bilan (Quadruple Bottom Line) qui considère les capacités numériques et la maturité des SI comme un levier de durabilité (Watanabe *et al.,* 2017). Cette perspective reste toutefois peu développée dans la littérature et ouvre des pistes intéressantes pour la recherche en SI. Plusieurs travaux mobilisent des indicateurs agrégés sans exploiter pleinement la structure des processus sous-jacents. Cela atténue l'intérêt de la fouille de processus, en rapprochant certaines approches de méthodes analytiques plus classiques, centrées sur des agrégations de données plutôt que sur l'analyse des dynamiques d'exécution.

5.4. RQ4 – Enjeux sociaux et éthiques

L'analyse met en évidence un décalage important entre l'importance théorique des enjeux sociaux et éthiques associés à l'application de la fouille de processus à la durabilité et leur prise en compte effective dans les travaux étudiés.

Les questions d'équité, de confidentialité et de transparence sont principalement abordées dans des travaux conceptuels ou méthodologiques sur la Green Data Science et la fouille de processus responsable (van der Aalst, 2016). En revanche, elles sont rarement intégrées de manière explicite dans les cadres analytiques appliqués à la durabilité. La majorité des études se concentre sur l'amélioration des performances environnementales ou économiques, sans analyser les effets potentiels des décisions issues de la fouille sur les individus ou les groupes concernés. L'actionnabilité des résultats constitue un autre enjeu clé. Si certains travaux proposent des mécanismes explicites reliant les analyses de fouille de processus à des décisions opérationnelles ou stratégiques (Kroeger *et al.,* 2024; Kumbhar *et al.*, 2022; Park et van der Aalst, 2022), d'autres se limitent à des analyses descriptives sans lien clair avec les processus décisionnels. Cela limite l'impact réel des approches sur la durabilité des SI.

Le Tableau 3 synthétise les principaux enjeux sociaux et éthiques associés à l'application de la fouille à la durabilité et leur prise en compte dans les travaux analysés. Il met en évidence un décalage entre l'identification conceptuelle de ces enjeux et leur opérationnalisation dans les SI décisionnels. L'analyse révèle que les dimensions sociales et éthiques restent largement sous-explorées dans les applications de la fouille à la durabilité. Leur intégration systématique apparaît comme un axe de recherche prioritaire pour concevoir des SI décisionnels capables de soutenir une durabilité à la fois opérationnelle, stratégique et responsable.

Tableau 3 – Enjeux sociaux et éthiques

Enjeu social / éthique	**Enjeux**	**Manifestation dans la fouille de processus**	**Prise en compte dans les travaux**
Equité	Risque de biais ou de discrimination	- découverte de processus biaisée	- largement absente des cadres appliqués

(van der Aalst, 2016)	induits par l'analyse des données de processus	- décisions automatisées défavorables à certains groupes	- traitée plutôt au niveau conceptuel
Confidentialité (van der Aalst, 2016)	Protection des données sensibles issues des SI et des logs de processus	- réidentification possible via horodatages - identifiants ou corrélations	- peu explicitée dans les études de durabilité - supposée mais rarement opérationnalisée
Transparence (Watanabe *et al.,* 2017), (Rai et Daniels, 2016), (Kroeger *et al.,* 2024), (Gribaudo et Manini, 2020)	Compréhensibilité et traçabilité des analyses et décisions issues de la fouille de processus	- modèles interprétables - diagnostics explicites - traçabilité des décisions	Relativement bien prise en compte dans les cadres industriels et de simulation
Actionnabilité (Kroeger *et al.,* 2024), (Kumbhar *et al.,* 2022), (Park et van der Aalst, 2022)	Capacité à transformer les résultats analytiques en décisions concrètes	- lien explicite entre indicateurs - contraintes et actions	- variable selon les travaux - formalisée dans les approches orientées décision

6. Discussion

L'analyse des travaux met en évidence le potentiel de la fouille de processus comme levier structurant pour l'aide à la décision en matière de durabilité, mais elle révèle surtout des différences fortes dans les choix de conception des systèmes analysés. Au-delà du constat désormais classique du déséquilibre en faveur des dimensions environnementales, les résultats montrent que ce déséquilibre est étroitement lié aux types de données mobilisées, aux cadres de modélisation retenus et aux formes d'aide à la décision effectivement proposées. Cette lecture permet de dépasser une analyse thématique pour mettre en évidence des mécanismes structurels propres aux SI décisionnels.

Nous pouvons noter la forte hétérogénéité des approches, tant sur les sources de données que sur les cadres de modélisation et les systèmes d'indicateurs mobilisés. Cela reflète la pluralité des contextes d'application, mais complique l'identification de cadres analytiques génériques et réutilisables pour la conception de SI décisionnels durables. Le cloisonnement entre approches organisationnelles et approches industrielles limite souvent la capacité à relier les décisions stratégiques aux impacts opérationnels réels. Les travaux proposant des cadres intégrés apparaissent prometteurs, mais restent encore minoritaires et coûteux à mettre en œuvre. La Figure 3 propose une synthèse des principaux apports et limites des approches analysées.

Un second point de discussion concerne la centralité des indicateurs environnementaux dans les applications de la fouille à la durabilité. Les indicateurs sociaux sont rarement intégrés de manière opérationnelle, en raison de leur caractère qualitatif, contextuel et difficilement mesurable à partir de données de processus.

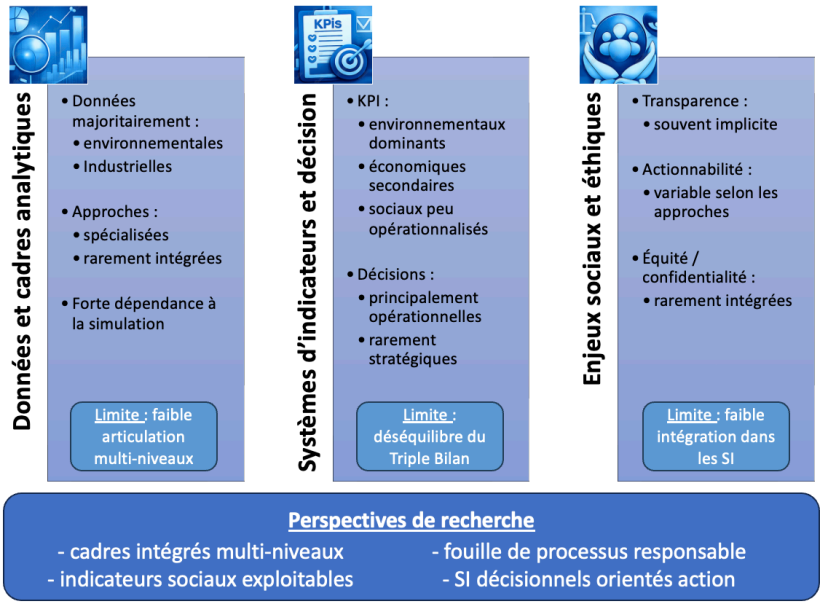


Figure 3 – Synthèse des apports et limites des approches de fouille de processus pour la durabilité.

Cette situation révèle une limite structurelle des approches actuelles, qui tendent à privilégier les dimensions de la durabilité les plus facilement quantifiables.

Par ailleurs, la discussion met en lumière le rôle encore marginal accordé aux enjeux éthiques et sociaux de la fouille de processus, notamment en termes d'équité, de confidentialité et de transparence. Bien que ces questions soient identifiées dans la littérature méthodologique, elles restent peu intégrées dans les cadres appliqués à la durabilité. Cette lacune est problématique pour la recherche en systèmes d'information, dans la mesure où l'aide à la décision fondée sur des données de processus peut produire des effets différenciés sur les acteurs et renforcer des asymétries existantes. L'intégration explicite de ces dimensions constitue ainsi un enjeu majeur pour le développement de SI réellement durables.

Un point central mis en évidence par cette analyse concerne la sous-exploitation des capacités propres à la fouille de processus. Alors que celle-ci permet d'analyser les processus réels à partir de leurs traces, notamment en termes de variantes d'exécution, de séquences d'activités ou de dynamiques temporelles, plusieurs travaux se limitent à des analyses agrégées. Cette situation réduit la valeur ajoutée spécifique de la fouille de processus et interroge son positionnement par rapport à des approches analytiques plus traditionnelles.

Enfin, l'analyse souligne l'importance de l'actionnabilité des résultats de la fouille de processus. Les approches les plus pertinentes pour l'aide à la décision sont celles qui établissent un lien explicite entre les analyses produites et les choix organisationnels ou opérationnels. À l'inverse, les travaux se limitant à des analyses descriptives peinent à démontrer leur valeur pour le pilotage de la durabilité. La fouille ne doit pas être envisagée uniquement comme un outil analytique, mais comme un composant des SI décisionnels.

Ce travail montre un décalage entre le potentiel théorique de la fouille de processus pour soutenir la durabilité et la maturité des cadres effectivement proposés. Ce décalage ouvre plusieurs perspectives de recherche, autour de la conception de cadres intégrés, de l'opérationnalisation des dimensions sociales et de l'intégration de principes éthiques dès la phase de conception des SI. Ces résultats soulignent également le rôle central de l'architecture des systèmes d'information dans la capacité à intégrer, transformer et exploiter les données de processus pour la durabilité.

7. Conclusion

Cet article propose une revue analytique structurée des travaux mobilisant la fouille de processus au service de la durabilité et de l'aide à la décision dans les SI. Nous proposons une grille d'analyse opérationnelle permettant de comparer les approches existantes et de guider la conception de systèmes d'information décisionnels orientés durabilité. Cela permet de structurer la littérature autour de quatre axes principaux : la constitution des données, les cadres de modélisation et d'évaluation, les systèmes d'indicateurs et les enjeux sociaux et éthiques. Les résultats montrent que la fouille de processus constitue un levier pertinent pour relier l'exécution réelle des processus aux objectifs de durabilité, en particulier dans les contextes industriels et organisationnels complexes. Toutefois, la littérature actuelle demeure fragmentée, avec une prédominance des approches environnementales et une prise en compte encore limitée des dimensions sociales, éthiques et technologiques de la durabilité. Cette situation limite la capacité des SI à soutenir une prise de décision véritablement intégrée et responsable.

Ce travail invite à repenser la fouille de processus non seulement comme un outil d'analyse des processus, mais comme un composant structurant de SI décisionnels durables. Il met en évidence la nécessité d'intégrer dès la conception des SI les choix relatifs aux données, aux indicateurs et aux principes éthiques, afin de garantir une aide à la décision à la fois actionnable et responsable. Ces résultats fournissent des points d'appui concrets pour la recherche en SI souhaitant concevoir des cadres analytiques réutilisables et alignés avec les enjeux contemporains de la durabilité.

Bibliographie

Acerbi F., P. W., Q. (2022). Fostering Circular Manufacturing through the integration of Genetic Algorithm and Process Mining. *Advances in Production Management Systems. Smart Manufacturing and Logistics Systems: Turning Ideas into Action*.

van der Aalst W. M. (2016). Green Data Science – Using Big Data in an "Environmentally Friendly" Manner. *International Conference on Enterprise Information Systems*.

Bauer J. (2024). *How process mining can help find a sustainability sweet spot*. Interview CELONIS, janvier 2024.

Docekalová M. P., D. M., A. K. (2017). Evaluations of corporate sustainability indicators based on fuzzy similarity graphs. *Ecological Indicators*, Elsevier.

Gribaudo M. , Manini D. (2020). Circular Economy: A Performance Evaluation Perspective. *ACM International Conference Proceedings*.

Graves N., K., W., van der Aalst W. M. (2023). ReThink Your Processes! A Review of Process Mining for Sustainability. *9th International Conference on ICT for Sustainability (ICT4S 2023)*, Rennes.

Joas A., Gierlich.-Joas M.. Bahr C., Bauer J. (2024). Towards Leveraging Process Mining for Sustainability – An Analysis of Challenges and Potential Solutions. *Business Process Management Forum*.

Kumbhar M., Ng A. HC., Bandaru S. (2022). Bottleneck Detection Through Data Integration, Process Mining and Factory Physics-Based Analytics. *Advances in Transdisciplinary Engineering*.

Kroeger S., Streibel L., Jordan P. Klages B. (2024). Sustainability assessment of production networks using simulation-data-based process mining. *Procedia Computer Science*.

Rai S., Daniels M. (2016). An event-log analysis and simulation-based approach for quantifying sustainability metrics in production facilities. *Winter Simulation Conference*.

Safitri, L. N., Sarno R., Budiawati G.I. (2018). *Improving Business Process by Evaluating Enterprise Sustainability Indicators Using Fuzzy Rule-Based Classification*. *International Seminar on Application for Technology of Information and Communication.*

van Dongen B. F., van der Aalst W. M., L. W. (2009). Process Mining: Overview and Outlook of Petri Net Discovery Algorithms. *Transactions on Petri Nets and Other Models*

Park G., van der Aalst W. M. (2022). Action-oriented process mining: bridging the gap between insights and actions. *Progress in Artificial Intelligence*.

Rabbanee, F. K., Hassan K., Taufique K.M.R., Shafiullah G.M., Dewan A. (2023). *Sustainability of Sustainable Business Practices: Challenges and Innovations*. *Sustainability*.

United Nations (2023). *Global Sustainable Development Report (GSDR)*.

United Nations (1992). *United Nations Conference on Environment and Development, Rio de Janeiro, Brazil, 3–14 June 1992*.

Union européenne (2024). *Directive (UE) 2022/2464 du Parlement européen et du Conseil du 14 décembre 2022 relative au reporting de durabilité des entreprises*.

Watanabe E. Marinho Da Silva R., Tsuzuki M., Junqueira F., Santos Filho D.J., Miygi P.E. (2017). Assessment of Sustainability for Production Control Based on Petri Net and Cyber-Physical Cloud System. *IFAC PapersOnLine*, Elsevier.

Apprentissage profond pour la caractérisation des séries temporelles de consommation électrique

Adrien Petralia[1 2]

1. Université Paris Cité, LIPADE
45 Rue des Saints-Pères, F-75006 Paris, France
2. EDF Lab Paris-Saclay, SEQUOIA
7 Bd Gaspard Monge, 91120 Palaiseau
adrien.petralia@gmail.com

RESUME. *Cet article propose une synthèse de la thèse de doctorat intitulée "Apprentissage profond pour la caractérisation des séries temporelles de consommation électrique", préparée par Adrien Petralia à Université Paris Cité, en partenariat avec EDF R&D, sous la direction du Professeur Themis Palpanas, et soutenue le 7 mai 2025.*

MOTS-CLÉS : *Apprentissage profond ; séries temporelles ; surveillance non intrusive de la charge .*

1. Introduction

La transition vers des systèmes énergétiques bas carbone impose une plus grande flexibilité du réseau électrique. Le développement des énergies renouvelables, en particulier solaire et éolienne, réduit la dépendance aux combustibles fossiles, mais introduit une variabilité importante de la production. Il devient alors nécessaire de mieux équilibrer l'offre et la demande, en améliorant la compréhension des usages et l'accompagnement des consommateurs.

Dans ce contexte, les compteurs communicants constituent une source d'information essentielle. Ils fournissent des séries temporelles de puissance totale consommée et ouvrent la voie à de nouveaux services de retour d'information, de conseil et de pilotage

énergétique. Cependant, leur exploitation reste difficile. Les données collectées sont agrégées, combinant les usages de plusieurs appareils, et généralement échantillonnées à très basse fréquence, typiquement toutes les 30 minutes en France et au Royaume-Uni. Cette faible résolution atténue fortement les signatures caractéristiques des appareils. À cela s'ajoutent la rareté des données fortement annotées, la variabilité des longueurs de séries temporelles et le volume massif des données industrielles. Cette thèse propose des approches fondées sur l'apprentissage profond pour analyser ces courbes de charge et caractériser les usages électriques domestiques.

2. Verrous scientifiques

Un objectif central est la surveillance non intrusive de la charge, ou *Non-Intrusive Load Monitoring* (NILM), qui consiste à estimer les usages individuels des appareils à partir de la consommation totale du foyer. La majorité des approches existantes s'appuie toutefois sur des signaux acquis à haute fréquence, où les transitions électriques et les signatures d'appareils sont mieux préservées. Ces hypothèses sont peu compatibles avec les données réellement disponibles chez les fournisseurs d'électricité, souvent mesurées à une fréquence beaucoup plus faible.

Un second verrou concerne les annotations. Les méthodes de désagrégation fondées sur l'apprentissage profond requièrent souvent des mesures individuelles de consommation pour chaque appareil, obtenues au moyen de capteurs spécifiques. Or ces données sont coûteuses à collecter, difficiles à produire à grande échelle et rarement disponibles en contexte industriel. Les fournisseurs disposent plus fréquemment d'informations faibles, par exemple la présence ou l'absence d'un appareil dans un foyer, sans indication précise sur ses périodes d'utilisation. La thèse s'attaque ainsi à trois défis complémentaires : détecter des appareils à partir de séries très basse fréquence, exploiter des données non étiquetées ou faiblement annotées, et concevoir des modèles précis et efficients pour de grandes bases de données.

3. Contributions

La première contribution étudie la détection d'appareils à très basse fréquence. Le problème est formulé comme une tâche de classification de séries temporelles visant à déterminer si un foyer possède un équipement donné. Un benchmark mené sur cinq jeux de données réels issus de compteurs intelligents montre que plusieurs appareils peuvent être détectés avec précision malgré un pas de 30 minutes. Les méthodes d'apprentissage profond, en particulier convolutives, surpassent nettement les approches traditionnelles lorsque les volumes de données augmentent.

La deuxième contribution introduit ADF et TransApp, un cadre de détection adapté aux séries longues, nombreuses et de longueur variable. L'*Appliance Detection Framework* segmente les courbes de charge en sous-séquences fixes, puis fusionne les prédictions pour produire une décision robuste à l'échelle du foyer. TransApp repose

sur une architecture Transformer pré-entraînée de manière auto-supervisée sur des séries non étiquetées, puis ajustée sur des données annotées. Cette stratégie exploite les grands volumes de données brutes disponibles chez les fournisseurs d'électricité tout en limitant la dépendance aux annotations. Les évaluations sur deux jeux de données réels montrent des gains en précision et en passage à l'échelle.

La troisième contribution concerne la localisation des périodes d'activation à partir d'étiquettes faibles. Nous proposons CamAL, une approche combinant classification de séries temporelles et explicabilité. CamAL n'utilise que des informations de présence ou d'absence d'appareil dans un foyer. Lorsqu'un appareil est détecté, des cartes d'activation de classe sont extraites d'un ensemble de classificateurs convolutifs, puis agrégées pour identifier les zones contributives du signal. CamAL estime ainsi les périodes probables d'activation sans annotations temporelles détaillées, avec des performances comparables à celles de méthodes fortement supervisées, tout en réduisant jusqu'à trois ordres de grandeur le besoin en annotations.

Ces travaux ont ensuite été intégrés dans DeviceScope, une application interactive destinée aux utilisateurs non experts. DeviceScope permet d'explorer les séries de consommation, de détecter la présence d'appareils et de visualiser leurs signatures d'activation. L'outil relie ainsi les modèles d'apprentissage profond à des usages concrets d'interprétation, pour les consommateurs comme pour les fournisseurs d'énergie.

Enfin, la thèse propose NILMFormer, une architecture Transformer séquence-à-séquence dédiée à la désagrégation de charge et adaptée à la non-stationnarité des séries de consommation. Les sous-séquences extraites d'une courbe de charge présentent en effet des variations statistiques importantes liées à l'usage intermittent des appareils. NILMFormer stationnarise les entrées par retrait des statistiques locales, réinjecte ces informations dans le Transformer, puis les utilise pour dénormaliser les prédictions. Il introduit également TimeRPE, un encodage positionnel fondé sur les informations temporelles. Les évaluations sur quatre jeux de données réels montrent que NILMFormer surpasse les méthodes séquence-à-séquence existantes en précision et en passage à l'échelle, y compris pour l'estimation quotidienne de la consommation des appareils.

4. Conclusion

Cette thèse montre que les courbes de charge issues des compteurs communicants permettent de caractériser les usages électriques domestiques, malgré leur faible fréquence, leur nature agrégée et la rareté des annotations. Les contributions couvrent la détection d'appareils, la localisation de leurs activations, leur visualisation interactive et l'estimation de leur consommation, en combinant apprentissage auto-supervisé, apprentissage faiblement supervisé, explicabilité et architectures Transformer.

Ces travaux ont donné lieu à plusieurs dépôts de brevets et ont été intégrés dans des solutions industrielles à grande échelle. Ils ouvrent ainsi la voie à des services énergétiques plus précis, capables d'accompagner les consommateurs et les fournisseurs dans la transition énergétique.