



**Titre de la thèse/Thesis title :** Approche hybride pour la détection d'anomalies basée sur les ontologies et l'apprentissage automatique : Application aux données environnementales

**Laboratoire d'accueil / Host Laboratory :** LIB : Laboratoire d'Informatique de Bourgogne

**Spécialité du doctorat préparé/Speciality :** Informatique

**Mots-clés / Keywords :** Anomalie, environnement, pollution, apprentissage automatique, représentation et extraction des connaissances.

## Descriptif détaillé de la thèse

La représentation et l'extraction de connaissances jouent un rôle crucial en analyse de données, en particulier dans le domaine de la surveillance environnementale et de la détection d'anomalies liées aux polluants. Il est impératif de développer des méthodes efficaces pour filtrer et extraire des informations pertinentes adaptées aux besoins spécifiques des utilisateurs, car elles transforment des données brutes en informations exploitables et significatives.

Ces techniques peuvent inclure l'utilisation d'approches d'apprentissage automatique pour identifier des tendances, des anomalies ou des modèles dans de vastes ensembles de données environnementales. Le traitement du langage naturel (TALN) joue également un rôle clé, permettant d'analyser et d'interpréter des données textuelles, telles que des rapports de recherche, des articles et des documents réglementaires, afin d'extraire des informations pertinentes sur les polluants et leurs effets. En outre, les approches basées sur la sémantique, incluant le développement d'ontologies et de modèles de données, sont indispensables pour modéliser les connaissances et faciliter le raisonnement.

Dans ce contexte, les graphes de connaissances apparaissent comme une solution puissante pour représenter de manière explicite les entités environnementales, leurs propriétés et leurs relations. Ils permettent de relier les données hétérogènes issues de sources variées, d'en faciliter l'interprétation et d'activer des mécanismes d'inférence pour enrichir la compréhension des phénomènes complexes. Leur construction repose en grande partie sur les ontologies, qui fournissent un cadre formel pour structurer les connaissances du domaine. Ces graphes, interrogeables permettent la visualisation des interactions entre pollu-

ants, sources, milieux impactés, et effets potentiels sur la santé ou les écosystèmes.

Ces dernières années, les ontologies sont devenues plus populaires pour exprimer des connaissances sémantiques lisibles par les machines. De nombreuses études se sont concentrées sur la construction d'ontologies dans le but de créer des représentations structurées des connaissances pour diverses applications [1, 2]. Une fois l'ontologie construite, deux processus distincts mais liés dans le domaine de l'extraction et la gestion de connaissances et de la modélisation sémantique sont l'enrichissement et le peuplement d'ontologies. Des chercheurs ont exploré la construction d'ontologies dans le domaine de la pollution, comme Oprea qui décrit une ontologie représentant les concepts de la pollution de l'air [8]. Ces ontologies définissent des concepts de haut niveau qui peuvent être utilisés dans la construction d'une ontologie de domaine spécifique aux polluants dans l'environnement.

Plusieurs travaux récents ont étudié la comparaison de modèles d'apprentissage automatique en fonction des données et du volume. Les approches basées sur l'apprentissage supervisé sont utilisées dans les problèmes de détection d'anomalies pour prédire les transactions suspectes [10]. Les approches basées sur l'apprentissage non supervisé sont employées pour la détection de transactions suspectes, en regroupant des comportements similaires. Les approches basées sur l'apprentissage semi-supervisé combinent les techniques d'apprentissage non supervisé pour labelliser les données avec les classes générées par le modèle, puis ces données labellisées alimentent un algorithme d'apprentissage supervisé [5].

Les modèles d'apprentissage profond complémentaires ouvrent de nouvelles possibilités [3, 12, 4]. Par rapport aux méthodes traditionnelles d'apprentissage automatique qui nécessitent des caractéristiques conçues manuellement et pour lesquelles la proximité spatiale et les corrélations temporelles à long terme des données sont généralement complexes à capturer. De plus, les modèles d'apprentissage profond fonctionnent mieux avec de plus grandes quantités de données. Ces modèles peuvent également approximer des fonctions non linéaires complexes et s'adapter à n'importe quelles courbes tant qu'ils ont suffisamment de couches et de neurones. Par exemple, avec l'opération de convolution multicouche dans les CNN (réseaux neuronaux convolutionnels) et la structure récurrente des RNN (réseaux neuronaux récurrents), la proximité spatiale et les corrélations temporelles dans les données peuvent être apprises automatiquement et efficacement à partir des données brutes. La principale difficulté réside dans le choix multiple des types d'architectures de réseau à expérimenter, tels que les auto-encodeurs profonds [9, 7, 6] et les transformers (modèles qui utilisent des mécanismes auto-attentionnels) [11]. Plus récemment, les transformers ont surpassé la littérature dans la modélisation des dépendances temporelles à long terme.

## Les verrous scientifiques

La présence de divers polluants dans l'environnement pose des défis majeurs et complexes, notamment en termes de qualité de l'air et de l'eau, santé des écosystèmes et impacts sur la santé humaine. Les anomalies environnementales se réfèrent à des phénomènes qui dévient de la norme, tels que des concentrations inhabituellement élevées de substances nocives, des changements inattendus dans les écosystèmes ou des comportements toxiques. La détection de ces anomalies présente des difficultés méthodologiques particulières.

- La complexité des écosystèmes et des réseaux de pollution, impliquant de nombreux facteurs et interactions, pose un défi pour l'analyse des données et l'extraction de connaissances. Cette complexité nécessite des méthodes d'analyse capables de comprendre les interactions et les impacts environnementaux.
- Le manque de données de référence : Pour identifier efficacement les anomalies environnementales, il est essentiel de disposer de données de référence fiables. Le manque de données environnementales publiques et détaillées limite les possibilités d'expérimentation et de validation. L'objectif est d'obtenir des données de référence à partir de diverses sources.
- L'évolution des sources de pollution : Les sources de pollution évoluent sur diverses échelles de temps. À court terme, l'utilisation de certains produits chimiques varie en fonction de la période de l'année. À plus long terme, nous observons l'introduction de nouveaux composés chimiques ayant un impact significatif sur l'environnement et la santé humaine. L'objectif est de développer des modèles capables de prédire et de suivre ces évolutions, afin d'anticiper les impacts environnementaux et sanitaires.

## Méthodologie

Ce sujet de thèse mobilise différents domaines d'expertise, notamment les sciences de données, la chimie des polluants et l'étude des impacts environnementaux. La collaboration interdisciplinaire entre ces domaines est essentielle pour approfondir notre compréhension et développer des solutions intégrées et efficaces.

### 1. Construction d'une ontologie et d'un graphe de connaissances des polluants :

- (a) *Gestion de la multiplicité des sources* : L'objectif est d'extraire automatiquement des entités et des relations à partir de documents scientifiques, réglementaires ou issus de bases de données environnementales, puis de les intégrer dans une ontologie et un graphe de connaissances interopérable.

- (b) *Structuration et alignement* : Il s’agira de concevoir des mécanismes d’enrichissement et d’alignement avec des graphes existants, pour favoriser l’interopérabilité et la réutilisabilité des connaissances.
- (c) *Validation et évaluation* : La validation des graphes construits passera par des scénarios d’usage (requêtes SPARQL, inférences, etc.) et des comparaisons avec des connaissances expertes.

## 2. Analyse de données et détection d’anomalies guidée par les graphes de connaissances :

- (a) Intégration de graphes de connaissances dans les processus de fouille de données, pour guider la sélection des caractéristiques ou contraindre les modèles.
- (b) Développement d’approches hybrides combinant apprentissage automatique et raisonnement sémantique, afin d’améliorer la détection d’anomalies et d’interpréter les résultats à la lumière des connaissances disponibles.

## Direction de la thèse / Thesis Supervisor

- **Lylia Abrouk** (40%) lylia.abrouk@u-bourgogne.fr

## Encadrement de la thèse

### Co-directeurs

- **Nadine Cullot** (30%) Co-directrice

### Co-encadrants

- **Claire Bourgeois-Republique** (30%) Co-encadrante

## References

- [1] Lylia Abrouk, Hamza Chergui, and Hamid Ahaggach. Ontofic : an ontology for financial fraud detection and customer behavior modeling. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2023, Kusadasi, Turkey, November 6-9, 2023*, pages 509–513. ACM, 2023.
- [2] Hamid Ahaggach, Lylia Abrouk, and Eric Lebon. Information extraction from automotive reports for ontology population. pages 1–30, 2024.
- [3] M. M. TL Alam and BA. A survey on spatio-temporal data analytics systems. *ACM Comput. Surv.*, 2022.

- [4] R. S. IM Aparna. Spatio-temporal data clustering using deep learning: A review. In *IEEE Int. Conf. on Evolving and Adaptive Intelligent Systems*, 2022.
- [5] Fabrizio Carcillo, Yann-Aël Le Borgne, Olivier Caelen, Yacine Kessaci, Frédéric Oblé, and Gianluca Bontempi. Combining unsupervised and supervised learning in credit card fraud detection. *Information Sciences*, 557:317–331, 2021.
- [6] Q. ZJ Ma, LS, and CG. Learning representations for time series clustering. In *Proc. of 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019.
- [7] N. S. SS Madiraju and FDea2. Deep temporal clustering: Fully unsupervised learning of time-domain features. *arXiv*, 2018.
- [8] Mihaela M. Oprea. Air pollution onto: An ontology for air pollution analysis and control. In *Artificial Intelligence Applications and Innovations III*, pages 135–143. Springer US, 2009.
- [9] T. KK Thinsungnoen and KN. Deep autoencoder networks optimized with genetic algorithms for efficient ecg clustering. *Int. Journal of Machine Learning and Computing*, 8, 2018.
- [10] Dejan Varmedja, Mirjana Karanovic, Srdjan Sladojevic, Marko Arsenovic, and Andras Anderla. Credit card fraud detection - machine learning methods. In *2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH)*, pages 1–5, 2019.
- [11] A. SN Vaswani and PN. Attention is all you need. In *Proc. of 30th Conf. on Neural Information Processing Systems*, 2017.
- [12] S. CJ Wang and YPS2. Deep learning for spatio-temporal data mining: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 34, 2022.