

## PROPOSITION D'UN SUJET DE THESE DANS LE CADRE D'UNE CONVENTION CIFRE

### Modèle générique de métadonnées centré qualité pour les Data Lake : Application aux données de santé

Laboratoire d'Accueil : LIAS EA 6315 Equipe : Ingénierie des Données et Des modèles Partenaire Industriel : TRIMANE Directeur de Thèse : Allel HADJALI, <a href="mailto:allel.hadjali@ensma.fr">allel.hadjali@ensma.fr</a>	Codirection : Fatma Abdelhedi (Directrice, CBI <sup>2</sup> ) Slimane Hammoudi (Prof ESEO, HdR) ESEO, Angers. <a href="mailto:slimane.hammoudi@eseo.fr">slimane.hammoudi@eseo.fr</a>	Partenaire Industriel : TRIMANE Responsable : Fatma Abdelhedi (Directrice, CBI <sup>2</sup> ) 102 Terrasse Boieldieu, 92800 Puteaux <a href="mailto:fatma.abdelhedi@trimane.fr">fatma.abdelhedi@trimane.fr</a>
--	--	---

#### 1. Contexte et problématiques

Cette thèse se situe dans le domaine de la gestion et l'analyse des données massives supportées par des Data Lake (lacs de données) avec des applications aux données de santé. Au cours de la dernière décennie, le concept de lac de données a émergé comme un prolongement des entrepôts de données pour le stockage et l'analyse des données massives. Le lac de données contient des ensembles de données diversifiés, depuis les feuilles des calculs jusqu'aux bases de données NoSQL *schemaless* en passant par la sauvegarde de données brutes sous forme de fichiers, de tableaux ou de séquences vidéos. Dans ce contexte, l'objectif de ce travail de thèse vise à apporter des solutions scientifiques aux problématiques de la détection d'entités et de liens ou bien de valeurs sémantiquement équivalentes dans les Data Lake et la caractérisation des possibles *homograph* (des valeurs similaires avec différentes sémantiques) [16], [17]. Ces problématiques sont cruciales pour d'une part réaliser un stockage et un requêtage consistants des données massives (multi-sources et multi-format), et pour d'autre part exploiter efficacement ces données lors des analyses. Les principales contributions visées dans cette thèse sont triples : Il s'agit dans un premier temps de bien identifier et caractériser l'ensemble des dimensions (structurelle, sémantique, terminologique, extensionnelle) permettant de découvrir ou prévenir dans un Data Lake soit des objets<sup>1</sup> représentant la même entité du monde réel ou bien une valeur avec plusieurs occurrences représentant un *homographe*. Dans un second temps et en se basant sur les dimensions identifiées, il s'agit de proposer et définir un modèle générique des métadonnées intégrant ces dimensions et permettant de gérer différents types de lacs de données ; ce modèle doit être adapté à des cas d'usages différents. Ainsi, un métamodèle générique permettant de décrire les métadonnées centré qualité sera proposé afin de couvrir d'une part toutes les caractéristiques et concepts des modèles de métadonnées proposés dans l'état de l'art et disposer d'autre part de métadonnées permettant d'assurer les critères de qualités souhaités pour les Data Lake sur la résolution d'entités et la désambiguation. Dans cette perspective, des techniques à base d'ontologies et de l'IA générative basée Machine Learning seront explorées pour la mise en œuvre des critères de qualité comme précisé dans l'article paru récemment dans AI Magazine<sup>2</sup> :

*A new data quality revolution is underway, powered by models that use generative AI and machine learning techniques such as ChatGPT.*

<sup>1</sup> Un objet peut se matérialiser par un tuple de valeurs, une table relationnelle ou un fichier physique (document de tableur, XML ou JSON, document textuel, liste de tweets, image, vidéo, etc.).

<sup>2</sup> <https://aimagazine.com/articles/generative-ai-and-ml-fuelling-a-revolution-in-data-quality>

La troisième contribution visée sera liée à la mise en œuvre, l'évaluation et la validation des concepts et techniques proposées. En se basant sur le métamodèle générique des métadonnées, une architecture fonctionnelle qui précise les composants nécessaires à la réalisation d'un lac de données sera proposée pour mettre en œuvre les critères de qualité développés dans cette thèse. Finalement, basée sur cette architecture, une implémentation du Lac de données avec des technologies existantes et justifiées sera réalisée avec un cas d'usage centré sur les données issues de la santé dans le cadre de la collaboration avec l'entreprise *Trimane*. Sachant que les lacs de données sont diversifiés tant sur les types que sur les formats de stockage et avec une qualité assurée, le cas d'usage devra permettre d'envisager un accès unifié à des professionnels de santé (personnels hospitaliers, de mutuelles, de cliniques). Il s'agit aussi de générer des niveaux d'analyse en temps réel inédits portant sur des données administratives ou des données cliniques.

## 2. Objectifs du projet de recherche

Nos travaux se situent à la croisée de trois domaines, les données massives brutes et hétérogènes (Data Lake), la qualité des données et l'entreposage des données. Ils visent à apporter de nouvelles solutions pour assurer une qualité à des données décisionnelles complexes et ainsi permettre leurs analyses de manière la plus fiable et optimale.

Les travaux menés dans le cadre de cette thèse **auront un triple objectif :**

- De bien identifier et caractériser l'ensemble des dimensions (structurelle, sémantique, terminologique, extensionnelle) permettant de découvrir ou prévenir dans un Data Lake soit des objets représentant la même entité du monde réel ou bien une valeur avec plusieurs occurrences représentant un homographe. Ces dimensions vont permettre de stocker en évitant la redondance et interroger de manière fiable les données d'un Data Lake.
- De proposer et définir un modèle générique des métadonnées intégrant les dimensions identifiées dans le 1<sup>er</sup> objectif et permettant de gérer différents types de lacs de données et adapté à des cas d'usages différents. Ainsi, un métamodèle générique de métadonnées centré qualité sera proposé afin de couvrir d'une part toutes les caractéristiques et concepts des modèles de métadonnées proposés dans l'état de l'art et disposer d'autre part de métadonnées permettant d'assurer les critères de qualités souhaités pour les Data Lake sur la redondance d'entités et la désambiguation.
- De proposer et mettre en œuvre une architecture fonctionnelle qui précise les composants nécessaires à la réalisation d'un lac de données intégrant les critères de qualité développés dans la thèse. Basée sur cette architecture, une implémentation du Lac de données avec des technologies existantes et justifiées sera réalisée avec un cas d'usage centré sur les données issues de la santé dans le cadre de la collaboration avec l'entreprise *Trimane*.

## 3. Candidature :



Envoyez par email et en PDF, les documents suivants :

- CV,
- lettre de motivation ciblée sur le sujet,
- au moins deux lettres de recommandation,
- relevés de notes (depuis le bac) + liste des enseignements suivis en M2 et en M1

Aux adresses mails suivantes :

[allel.hadjali@ensma.fr](mailto:allel.hadjali@ensma.fr);  
[fatma.abdelhedi@trimane.fr](mailto:fatma.abdelhedi@trimane.fr);  
[slimane.hammoudi@eseo.fr](mailto:slimane.hammoudi@eseo.fr)

## Bibliographie

- [1] F. Nargesian, E. Zhu, R. J. Miller, K. Q. Pu, and P. C. Arocena, "Data lake management: challenges and opportunities," *Proc. VLDB Endow.*, vol. 12, no. 12, pp. 1986–1989, Aug. 2019.
- [2] A. Tunjić, "The Automation of the Data Lake Ingestion Process from Various Sources," in 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), May 2019, pp. 1276–1281.
- [3] P. N. Sawadogo, T. Kibata, and J. Darmont, "Metadata Management for Textual Documents in Data Lakes," in 21st International Conference on Enterprise Information Systems (ICEIS 2019), Heraklion, Greece, May 2019, vol. 1, pp. 72–83.
- [4] R. Lima and E. Cruz, "Extraction and Multidimensional Analysis of Data from Unstructured Data Sources: A Case Study," presented at the 21st International Conference on Enterprise Information Systems, May 2020, pp. 190–199.
- [5] A. Beheshti, B. Benatallah, Q. Z. Sheng, and F. Schiliro, "Intelligent Knowledge Lakes: The Age of Artificial Intelligence and Big Data," in *Web Information Systems Engineering*, Singapore, 2020, pp. 24–34.
- [6] J. Yeung, S. Wong, A. Tam, and J. So, "Integrating Machine Learning Technology to Data Analytics for E-Commerce on Cloud," in 2019 Third World Conference on Smart Trends in Systems Security and Sustainability (WorldS4), Jul. 2019, pp. 105–109.
- [7] H. Cheng, H. Fang, and M. Ostendorf, "A Dynamic Speaker Model for Conversational Interactions," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, Jun. 2019, pp. 2772–2785.
- [8] S. Arsovski, H. Osipyan, M. I. Oladele, and A. D. Cheok, "Automatic knowledge extraction of any Chatbot from conversation," *Expert Syst. Appl.*, vol. 137, pp. 343–348, Dec. 2019. [9] Hadrien Diesbecq, Data fabric, data hub, data Lake...quelles différences ? Mars 2022. <https://www.yzr.ai/articles/data-fabric-data-hub-data-lake-quelles-differences/>
- [10] DPO Partagé, L'entrepôt de données de santé: un outil incontournable pour la gestion et l'analyse des données médicales, Mars 2023. <https://www.dpo-partage.fr/entrepot-de-donnees-de-sante/>
- [11] Health Data Hub : Des enjeux et des contraintes. 2023 <https://www.msconnect.fr/innovation-sante/esante/health-data-hub-des-enjeux-et-des-contraintes/>
- [12] Etienne Scholly, « De la modélisation des métadonnées à la conception d'un lac de données. Application à l'habitat social ». Thèse de Doctorat, Université de Lyon, Mai 2022.
- [13] « A new paradigm for managing data », MIT Technology Review Insights, 2023.
- [14] Othmane Azeroual , Meena Jha 2 , Anastasija Nikiforova , Kewei Sha , Mohammad Alsmirat and Sanjay JhaA. Record Linkage-Based Data Deduplication Framework with DataCleaner Extension. *Technol. Interact.* 2022, 6, 27. <https://doi.org/10.3390/mti6040027>
- [15] Riccardo CAPPUZZO. Deep Learning Models for Tabular Data Curation. Thèse de Doctorat, Université de Sorbonne, Avril 2022.
- [16] Renée J. Miller, Professor, Northeastern University, USA ; A Vision for Data Alignment and Integration in Data Lakes. *IEEE Big Data 2022 Osaka, Japan- Keynote Lecture*
- [17] Aristotelis Leventidis, Laura Di Rocco, Wolfgang Gatterbauer, Renée J. Miller, Mirek Riedewald: Domainet: Homograph Detection for Data Lake Disambiguation. *CoRR abs/2103.09940* (2021).
- [18] Rym Jemmali , Fatma Abdelhédi, Gilles Zurfluh: DLToDW: Transferring Relational and NoSQL Databases from a Data Lake. *SN Comput. Sci.* 3(5): 381 (2022)
- [19] Fatma Abdelhédi, Rym Jemmali , Gilles Zurfluh: Data Ingestion from a Data Lake: The Case of Document-oriented NoSQL Databases. *ICEIS (1) 2022: 226-233*
- [20] Yassine GUERMAZI: Résolution d'entités à base de transformeurs : application à la validation des noms et adresses d'entreprises. THÈSE DE DOCTORAT Soutenue à Aix-Marseille Université le 03 juillet 2023.