

Enrichissement de graphes de connaissances et modélisation de l'incertitude pour l'analyse des réseaux historiques

Encadrants : Cédric du MOUZA (dumouza@cnam.fr, lab. CEDRIC, CNAM Paris)
Stéphane LAMASSÉ (stephane.lamasse@univ-paris1.fr, lab. LAMOP, Univ. Panthéon Sorbonne)
Camelia CONSTANTIN (camelia.constantin@lip6.fr, LIP6-Équipe BD, Sorbonne Université)

Contexte et enjeux de la transition vers les graphes. Le passage des sources historiques manuscrites vers des formats numériques a longtemps été marqué par la dépendance aux bases de données relationnelles. Si le format tabulaire a permis de quantifier certains phénomènes et de soutenir l'histoire sérielle, il se révèle peu adapté pour représenter des parcours biographiques complexes ou des questionnements qui évoluent au fil du temps. L'adoption de graphes de connaissances (KG) marque ainsi un tournant méthodologique majeur, en offrant un modèle où l'information n'est plus conçue comme un enregistrement isolé, mais comme un réseau de relations dynamiques entre personnes, lieux, événements et documents au sein de corpus historiques. Pour l'historien, le graphe devient un véritable outil d'enquête, capable de faire émerger des structures de parenté, des réseaux de sociabilité et des configurations relationnelles qui restent difficiles à appréhender dans une organisation purement tabulaire.

Cependant, cette modélisation se heurte à la nature même du document historique : l'incomplétude des archives et la fragmentation des séries constituent une contrainte structurelle qui limite la complétude des graphes de connaissances construits à partir de ces sources. À cela s'ajoutent l'ambiguïté des noms de personnes, les homonymies, la variabilité orthographique et la présence d'identités partiellement attestées, qui rendent la désambiguïsation et le chaînage d'entités particulièrement délicats dans les corpus historiques. Le défi n'est plus seulement de stocker l'information, mais de représenter fidèlement le flou et l'incertitude [6] qui l'entourent sans trahir la rigueur de la critique historique.

La problématique de l'incertitude : une difficulté scientifique nouvelle. La problématique centrale de cette thèse réside dans l'extraction et la quantification de l'incertitude, une dimension souvent ignorée par les systèmes de gestion de données classiques qui privilégient un modèle déterministe et supposent des faits complets et certains. Dans un contexte médiéval, l'incertitude est omniprésente : elle affecte les propriétés d'un nœud, comme une date de naissance approximative ou un statut social mal attesté, mais aussi l'existence même d'une relation, par exemple une filiation simplement supposée ou discutée dans des sources divergentes. La difficulté est ici double et constitue une nouveauté scientifique dans le champ des graphes de connaissances appliqués aux sources historiques. D'une part, il faut pouvoir établir des stratégies de liage d'entités dans un environnement où les données sont instables, fragmentaires et parfois contradictoires. Comment affirmer que deux mentions de noms proches dans des documents différents désignent la même personne physique alors que leurs attributs, tels que les lieux ou les dates, sont partiels, bruités ou incompatibles, tout en contrôlant explicitement les taux d'erreur de liage ? D'autre part, la thèse devra résoudre le problème de l'agrégation : comment fusionner deux nœuds représentant potentiellement la même entité tout en mettant à jour, de manière cohérente, les scores d'incertitude associés aux faits et aux relations du graphe ? Il s'agit de traiter mathématiquement le renforcement, lorsque deux sources indépendantes concordent, ou au contraire la contradiction, en modélisant la confiance dans les triplets et en intégrant des mécanismes de fusion incertaine, tout en gardant une traçabilité fine des entités et des sources d'origine pour permettre à l'historien de remonter systématiquement à la source primaire [4].

L'apport de l'Intelligence Artificielle : du NLP aux GNN. L'Intelligence Artificielle constitue le levier technologique indispensable pour lever ces verrous en intervenant à chaque étape de la chaîne de traitement. Dans un premier temps, les modèles de traitement du langage naturel (NLP) et les grands modèles de langage (LLM) seront mobilisés pour extraire l'information tout en détectant les marqueurs linguistiques de l'incertitude, en s'appuyant sur les travaux de détection automatique des hedge cues et des segments spéculatifs dans les textes. Cette approche dépasse le simple repérage d'entités pour devenir une véritable évaluation de la fiabilité de l'information brute, où l'IA associe à chaque affirmation textuelle un score de certitude ou de spéculation afin de distinguer les faits établis des informations hypothétiques ou douteuses. Ensuite, l'IA appliquée aux graphes, et plus particulièrement les Graph Neural Networks (GNN), permettra de transformer le liage d'entités en une tâche d'apprentissage profond exploitant le contexte relationnel global du graphe plutôt que les seuls attributs locaux. Contrairement aux méthodes classiques, les GNN peuvent apprendre des représentations qui intègrent la position d'un individu dans le réseau social et la structure des relations qui l'entourent, facilitant ainsi la réconciliation de nœuds même lorsque leurs attributs textuels divergent ou sont incomplets. Enfin, l'apprentissage automatique sera utilisé pour l'inférence de connaissances, permettant à la fois de découvrir des relations manquantes et de propager les scores d'incertitude à travers le graphe, dans l'esprit des approches de knowledge graph completion.

Données. Le travail de recherche s'appuiera sur les données prosopographiques des bases Studium et Fasti, offrant un terrain d'expérimentation d'une richesse rare sur les élites universitaires et ecclésiastiques médiévales. Ces corpus ne sont pas seulement des réservoirs biographiques ; ils constituent des structures relationnelles complexes où le silence des sources

et les contradictions documentaires sont la norme plutôt que l'exception. En mobilisant ces données, l'enjeu sera de transcender le modèle déterministe traditionnel pour modéliser des graphes de connaissances intégrant la notion d'incertitude. Les bases Studium et Fasti recèlent en effet des attributs fragiles, tels que des dates de décès exprimées par des fourchettes incertaines ou des fonctions dont la chronologie se chevauche de manière incohérente, qui serviront de variables pour tester des algorithmes de liage d'entités sous contrainte d'incertitude.

Méthodologie : Extraction, Liage et Agrégation sous Incertitude. Le doctorant devra en premier lieu développer des méthodes innovantes fondées sur le traitement du langage naturel (NLP) et l'apprentissage profond pour extraire non seulement les entités nommées, mais aussi des indices de confiance et d'incertitude finement calibrés, en s'inspirant des approches de détection de spéculation et de modélisation probabiliste des affirmations textuelles. Ces scores ne dépendront pas uniquement de la clarté du texte, mais seront corrélés au contexte sémantique global et à une évaluation de la qualité des sources historiques, suivant les travaux qui intègrent des métriques de fiabilité contextuelles dans l'extraction d'informations incertaines. Cette étape est cruciale pour transformer une donnée textuelle brute en un objet probabiliste riche, capable d'être intégré dans la structure du graphe de connaissances, comme le proposent les pipelines d'extraction enrichis en incertitude pour des applications en KG.

Dans un second temps, les travaux porteront sur l'élaboration d'algorithmes de liage et d'agrégation spécifiquement conçus pour être « uncertainty-aware », en ligne avec les cadres récents de entity resolution probabiliste et de fusion sous incertitude. Plusieurs types d'approches sont envisagées pour relever ce défi. L'algorithmique de graphe couplée à l'IA, notamment à travers les Graph Neural Networks (GNN), sera mobilisée pour capturer la topologie du réseau et l'utiliser comme levier de réconciliation, en exploitant les représentations structurelles pour résoudre les ambiguïtés même en présence de données bruitées ou partielles. Parallèlement, d'autres approches d'IA comme l'apprentissage par métrique (metric learning) ou les modèles de bi-encodeurs seront explorées pour le liage d'entités, en adaptant des techniques qui génèrent des embeddings tenant compte de l'incertitude ou de la variabilité des sources.

La difficulté majeure, et l'un des verrous scientifiques de la thèse, résidera dans la nécessité d'adapter ces modèles, traditionnellement déterministes, pour qu'ils intègrent nativement l'incertitude, comme le soulignent les analyses des limites des approches classiques face à des données historiques fragmentaires. Il s'agira de proposer des fonctions de similarité avancées capables de traiter des valeurs floues ou des intervalles de confiance, et de définir des opérateurs d'agrégation aptes à gérer le renforcement ou l'atténuation de la confiance lors de la fusion de sources multiples. Ces opérateurs permettront de mettre à jour dynamiquement les connaissances du graphe de connaissances (KG), en répercutant chaque nouvelle information sur l'ensemble du réseau relationnel tout en préservant la traçabilité indispensable à l'analyse historique, conformément aux principes de provenance et de vérification probabiliste dans les KG.

Adéquation à l'institut. La méthodologie proposée est intrinsèquement interdisciplinaire, croisant l'expertise en algorithmique de graphes et IA du LIP6 [3, 2, 11], la modélisation et gestion de données du CEDRIC (projets Mastodons QUALHIS, ANR DAPHNÉ¹ et ANR LAURA² et [1, 7, 9, 10]) et la connaissance des sources et l'expertise historique pour les enquêtes et leur interprétation du LAMOP [1, 5, 8]. Ce projet s'inscrit pleinement dans l'ADN de SCAI en proposant une approche exploratoire située à la confluence de l'intelligence artificielle et des sciences humaines. En combinant des méthodes avancées de traitement du langage (LLM, NLP) pour l'extraction de l'incertitude avec des algorithmes d'IA sur graphes (GNN) et des structures de données sémantiques, ce travail répond directement aux défis de l'IA hybride portés par l'institut. L'originalité scientifique réside dans la capacité à transformer des problématiques historiques complexes en verrous technologiques pour l'informatique, faisant de l'interdisciplinarité non seulement un cadre d'application, mais un moteur d'innovation pour la conception de modèles de réconciliation d'entités résilients à l'imprécision des données. Outre des publications dans les deux communautés, nous envisageons une validation de l'applicabilité de la méthode à différents cas d'usage historiques au sein du Consortium Européen HÉLOÏSE-consortium d'historiens spécialisés en prosopographie.³

Au-delà de ce cadre, la réussite de ce projet pourrait servir de point de départ pour de nouvelles collaborations interdisciplinaires au sein de l'Alliance Sorbonne Université et de SCAI. Les modèles de gestion de l'incertitude et de liage d'entités développés pourraient ainsi être transposés à d'autres domaines partenaires : qu'il s'agisse de collaborations purement informatiques sur la robustesse des données avec l'Inria ou l'UTC, ou de nouveaux ponts vers les SHS et la santé. Les méthodologies de la thèse pourraient notamment trouver un écho au Muséum national d'Histoire naturelle (MNHN) pour la structuration de collections biologiques, ou encore à l'INSERM et l'AP-HP pour l'analyse de parcours de soins complexes où la réconciliation d'informations fragmentaires et incertaines est un enjeu important.

Références

- [1] J. Akoka, I. Comyn-Wattiau, S.Lamassé, and C. du Mouza. Modeling Historical Social Networks Databases. In *Proc. Intl. Conf. on System Sciences, (HICSS)*, pages 1–10, 2019.

¹<https://daphne-anr.huma-num.fr/>

²<https://anr.fr/Projet-ANR-25-CE38-0700>

³Héloïse : <https://heoise.hypotheses.org/>

- [2] Y. Bai, C. Constantin, and H. Naacke. Leiden-fusion partitioning method for effective distributed training of graph embeddings. In *ECML PKDD 2024*, volume 14947, pages 366–382. Springer, 2024.
- [3] C. Constantin, C. du Mouza, and Y. Li. A label-based edge partitioning for multi-layer graphs. In *52nd Hawaii International Conference on System Sciences, HICSS 2019, Grand Wailea, Maui, Hawaii, USA, January 8-11, 2019*, pages 1–10, 2019.
- [4] A. Djeddi, Y. Tzitzikas, et al. Uncertainty management in the construction of knowledge graphs : a survey. *Transactions on Graph Data and Knowledge*, 3(1) :1–40, 2024.
- [5] C. du Mouza, S. Lamassé, and J.-P. Genet. Uncertainty and prosopography : the case of the Studium Parisiense database (Université Paris 1 Panthéon-Sorbonne, France). In *IX Héloïse Workshop –European Network on Digital Academic History*, 2019.
- [6] W. M. Kouadri, J. Akoka, I. Comyn-Wattiau, and C. du Mouza. Uncertainty detection in historical databases. In *NLDB 2022*, page 73–85. Springer-Verlag, 2022.
- [7] W. M. Kouadri, J. Akoka, I. Comyn-Wattiau, and C. du Mouza. Uncertainty detection in historical databases. In *Proc. Intl. Conf. on Applications of Natural Language to Information Systems (NLDB)*,, pages 73–85, 2022.
- [8] T. Kouamé and S. Lamassé. The Quest for Origins. Academic Filiations in the Studium Parisiense Database. *Specimina Nova*, (13) :47–61, 2024.
- [9] M. Prieur, C. du Mouza, G. Gadek, and B. Grilhères. Evaluating and improving end-to-end systems for knowledge base population. In *ICAART*, pages 641–649, 2023.
- [10] M. Prieur, C. du Mouza, G. Gadek, and B. Grilhères. Shadowfax : Harnessing textual knowledge base population. In *SIGIR*, 2024.
- [11] H. Rahimi, H. Naacke, C. Constantin, and B. Amann. ANTM : aligned neural topic models for exploring evolving topics. *Trans. Large Scale Data Knowl. Centered Syst.*, pages 76–97, 2024.