GraphRAG Multimodal pour l'Interrogation Sémantique de Publications Scientifiques

Proposition de sujet de thèse

Equipe Bases de Données – LIP6 – Sorbonne Université, Paris

Résumé

Ce projet de thèse vise à concevoir un cadre GraphRAG multimodal pour l'interrogation sémantique de la littérature scientifique. Il s'agit d'unifier l'analyse du contenu (texte, figures, tableaux) et des métadonnées contextuelles (citations, benchmarks) pour modéliser les relations complexes entre fragments d'information. La méthodologie repose sur trois piliers : l'adaptation d'encodeurs multimodaux, l'alignement des données dans un espace latent commun, et la construction d'un graphe documentaire hétérogène augmenté par des ressources externes. Ces travaux visent à produire des outils de recherche d'information plus précis et explicables, favorisant ainsi la reproductibilité scientifique.

1 Contexte

Les publications scientifiques modernes sont intrinsèquement multimodales : elles combinent texte, figures et tableaux [11, 15], tout en s'inscrivant dans un réseau dense de références bibliographiques. Pour interroger ces archives, les systèmes de Génération Augmentée par la Recherche (RAG), fondés sur les Grands Modèles de Langage (LLM), remplacent progressivement les moteurs de recherche classiques grâce à leur capacité à représenter le contenu sémantique des articles et à inférer des relations implicites. Néanmoins, ces approches atteignent aujourd'hui leurs limites, notamment face à des requêtes exigeant une compréhension globale et relationnelle du domaine.

Cette limite devient particulièrement visible dès que l'on dépasse le document lui-même (souvent au format PDF) : les travaux scientifiques s'inscrivent désormais dans un écosystème empirique ouvert, lié à des artefacts internes (tableaux, figures, définitions, preuves, bibliographie) et externes (dépôts GitHub, articles cités, jeux de données de référence, plateformes de *benchmark*). Si le corps textuel fournit une vue narrative d'ensemble, ces ressources complémentaires sont cruciales pour une interprétation exhaustive et la reproductibilité des résultats [11]. Dans cette perspective, le GraphRAG (*Graph Retrieval-Augmented Generation*), fondé sur des graphes de connaissances ou de structure documentaire, modélise les relations explicites entre entités et concepts, offrant aux LLM un contexte plus riche, structuré et traçable [10].

Cependant, les architectures GraphRAG actuelles restent majoritairement limitées à l'analyse textuelle. Cette thèse propose de dépasser ces approches fragmentaires en développant un cadre de représentation unifié pour analyser et interroger les travaux scientifiques de manière globale, en exploitant les modalités internes et externes aux publications. En s'appuyant sur l'expertise du LIP6 en modélisation par graphes, l'enjeu est d'intégrer efficacement les *embeddings* visuels et tabulaires au cœur des systèmes GraphRAG, afin d'améliorer la réponse aux questions complexes et la recherche d'information scientifique multimodale [11, 15].

1.1 État de l'art

Actuellement, l'analyse automatique de la littérature scientifique s'appuie sur plusieurs familles d'approches :

— La majorité des systèmes convertissent encore les documents en texte brut, en s'appuyant sur des modèles de fondation performants (ex. : Nougat [3]). L'émergence des Modèles de Langage Visuels (VLM) spécialisés pour la science et des approches de rétro-ingénierie visuelle (ex. : DePlot [16], ChartQA [17]) permet désormais d'extraire des informations pertinentes à partir de figures

complexes (diagrammes, graphes de flux). Néanmoins, ces informations restent souvent silotées et déconnectées du contexte textuel et tabulaire global du document [14, 7, 19].
- Les archives scientifiques comme arXiv et OpenAlex [18] sont largement exploitées pour l'analyse de la dynamique scientifique [8]. Si ces archives permettent de tracer l'évolution des thématiques à une échelle macroscopique [21], elles traitent généralement les publications comme des « boîtes noires » textuelles ou sémantiques. Elles modélisent avec précision la topologie des citations (qui cite qui et quand), mais ignorent le contenu exact qui justifie réellement ces basculements de paradigmes.
- De plus en plus de conférences demandent aux auteurs de compléter leurs travaux par des références vers des dépôts GitHub contenant les ressources logicielles et les données nécessaires à la reproductibilité expérimentale. Des plateformes comme Kaggle ou *PapersWithCode* (fermée en 2025) ont visé à centraliser l'évaluation des modèles en reliant publications, implémentations (code source) et performances empiriques sur des jeux de données de référence (*benchmarks*). Ces dépôts restent toutefois sous-exploités faute d'outils d'analyse adaptés.

Ces approches ont chacune permis des avancées substantielles, mais leur juxtaposition met surtout en évidence la nécessité d'un cadre unificateur capable de relier ces sources d'information hétérogènes. L'essor récent des architectures RAG, popularisées par des assistants de recherche avancés fondés sur l'ancrage documentaire restrictif (*source-grounding*, à l'instar de NotebookLM), a prouvé l'efficacité de ces systèmes pour limiter les hallucinations des LLM sur des corpus personnels restreints [12]. Néanmoins, ces approches classiques s'appuient historiquement sur la recherche vectorielle de segments textuels isolés (*chunks*) [9]. Si elles excellent dans l'extraction factuelle intra-documentaire, elles échouent à modéliser la topologie macroscopique du savoir scientifique et à effectuer des raisonnements multi-sauts (*multi-hop reasoning*) à l'échelle d'une archive entière.

C'est pour franchir ce cap conceptuel que le GraphRAG a récemment émergé et démontré sa supériorité [6] en substituant les bases de données vectorielles plates par des graphes de connaissances (où les nœuds représentent des concepts, des auteurs ou des entités, et les arêtes leurs relations sémantiques). Toutefois, ces graphes sont aujourd'hui construits et interrogés de manière presque exclusivement textuelle. Ils ignorent les vecteurs d'information riches et structurés provenant des autres modalités, qui constituent pourtant le coeur de l'argumentation scientifique. Le verrou principal aujourd'hui n'est donc plus la simple capacité à extraire une modalité spécifique (table, figure, image), mais bien l'absence d'un système capable d'harmoniser les *embeddings* de toutes ces modalités dans un espace sémantique partagé pour alimenter ces architectures de raisonnement par graphes [11].

Les pistes de recherche à explorer pour atteindre ces objectifs sont multiples :

1. Construire un pipeline d'extraction documentaire fiable : segmentation article–figure–tableau, normalisation des métadonnées (légendes, sections, références croisées), puis création d'unités documentaires homogènes pour l'indexation.
2. Adapter des encodeurs spécialisés par modalité : comparer des variantes d'encodeurs texte, tableau et vision, avec des protocoles d'évaluation dédiés (qualité de représentation, robustesse au bruit OCR, sensibilité au domaine scientifique).
3. Apprendre un espace latent partagé : tester des stratégies contrastives, des projections guidées par la structure du document, et des mécanismes d'attention croisée afin de préserver à la fois la similarité sémantique et la spécificité de chaque modalité.
4. Concevoir un graphe documentaire multimodal : définir une typologie de nœuds et d'arêtes (citations, inclusion figure/section, dépendances méthodologiques, liens code–données), puis étudier l'impact de différents schémas de pondération sur la récupération de contexte.
5. Déployer et évaluer un GraphRAG multimodal : comparer l'approche à des baselines textuelles et hybrides sur des questions nécessitant justification explicite, raisonnement multi-hop et traçabilité fine des sources.

2 Problématique et objectifs scientifiques

La problématique centrale de cette thèse se formule ainsi : *Comment aligner et intégrer sémantiquement des représentations issues de modèles hétérogènes (texte, tableaux, figures) au sein d'un espace unifié, afin de démultiplier les capacités de raisonnement des LLM via une architecture GraphRAG multimodale ?*

Pour illustrer concrètement ce besoin d'intégration, l'architecture cible devra rendre les LLM capables de répondre à des requêtes scientifiques hautement contextuelles, telles que :

— « *Quelles sont les valeurs exactes des hyperparamètres (Tableau 2) responsables de la chute de performance illustrée dans la courbe de la Figure 4 et discutée à la Section 3 ?* »
— « *Le mécanisme d'attention décrit formellement dans le paragraphe 4.1 correspond-il fidèlement aux blocs logiques du diagramme d'architecture visuel de la Figure 1 ?* »
— « *Comment les résultats d'ablation de la méthode proposée (Tableau 1) se traduisent-ils visuellement dans les graphes de dispersion, et en quoi cela contredit-il les hypothèses de l'état de l'art mentionnées en introduction ?* »
— « *Comment les limites expérimentales identifiées dans la Figure 3 de cet article ont-elles été surmontées dans les architectures proposées par les articles qui le citent l'année suivante ?* »
— « *Comment les performances documentées dans les tableaux d'expérimentation montrent-elles le basculement progressif du paradigme des CNN vers les Transformers entre 2019 et 2023 dans cette sous-discipline ?* »

Pour atteindre ce niveau de raisonnement, les travaux s'articuleront autour de trois objectifs :

1. Identifier et adapter les meilleurs modèles d'encodage existants pour les structures complexes (tableaux, figures). Il s'agira de valider leur capacité à préserver l'information topologique et visuelle sur des tâches de référence, afin de garantir la qualité des *embeddings* en entrée du système [22, 5].
2. Faire de l'alignement inter-modalités un objet d'étude à part entière, fortement couplé à la modélisation en aval. Cet objectif vise à concevoir, comparer et évaluer différents paradigmes de fusion afin d'identifier l'architecture optimale en fonction des exigences topologiques des graphes cibles (ex. : relations spatiales pour un graphe de mise en page vs relations causales pour un graphe d'expérimentation) [15].
3. Modéliser le graphe documentaire intégrant ces modalités comme des nœuds enrichis, et déployer l'architecture GraphRAG en exploitant les mécanismes des graphes hétérogènes pour la contextualisation inter-modale. L'impact du système global sera mesuré sur des cas d'usage de compréhension scientifique complexe à l'aide de jeux de données de référence [10, 14].

3 Adéquation du sujet avec l'expertise de l'équipe encadrante

La réalisation de cette thèse s'appuiera sur un environnement scientifique en adéquation avec les verrous technologiques ciblés. Elle sera dirigée par Bernd Amann, et co-encadrée par Camélia Constantin et Rafael Angarita, au sein de l'équipe Bases de données du laboratoire LIP6.

Le sujet s'inscrit directement dans les compétences de l'équipe Bases de Données du LIP6 :

— analyse de graphes de citations et dynamique des sciences [13, 20, 21] ;
— représentation de données structurées (notamment tabulaires) [4, 5] ;
— ingénierie de graphes de connaissances à grande échelle et enrichissement sémantique [1, 2].

4 Résultats attendus et Impact

Les résultats attendus de cette thèse résident dans la création d'un cadre méthodologique original pour l'alignement sémantique de données hétérogènes, capable de transformer des archives scientifiques statiques en graphes de connaissances dynamiques et interconnectés. Concrètement, ces travaux visent à produire un pipeline d'extraction robuste et une architecture GraphRAG innovante permettant aux LLM d'effectuer des raisonnements multi-sauts complexes avec une précision accrue et une réduction significative des hallucinations. Au-delà de la performance technique, les retombées incluent le développement d'outils d'exploration capables de corréler explicitement des affirmations textuelles avec des preuves visuelles (figures) ou structurées (tableaux), offrant ainsi aux chercheurs une traçabilité fine des sources, une meilleure explicabilité des réponses et un soutien concret à la reproductibilité scientifique.

5 Contexte du poste

Modalités d'encadrement : La doctorante ou le doctorant sera encadré par deux membres de l'équipe Bases de Données du LIP6. Des réunions hebdomadaires de travail permettront d'assurer le pilotage scientifique et l'orientation méthodologique du travail.

Suivi de la formation doctorale : Le parcours de formation sera construit avec l'école doctorale, en articulation avec les besoins du projet : cours scientifiques avancés (apprentissage multimodal, graphes, LLM), formations transversales (éthique, science ouverte) et développement de compétences en communication scientifique.

Suivi de l'avancement des recherches : L'avancement fera l'objet d'un suivi structuré autour de jalons semestriels : revue de littérature, validation expérimentale, publications et intégration système. Les résultats seront discutés lors de points d'étape élargis à l'équipe, et consolidés dans des rapports intermédiaires afin de garantir la cohérence scientifique, la qualité méthodologique et la progression vers la soutenance.

6 Profil et compétences recherchées

Le poste s'adresse à une candidate ou un candidat motivé(e) par la recherche interdisciplinaire à l'interface entre traitement automatique des langues, vision par ordinateur et graphes de connaissances.

- Compétences scientifiques : bases solides en apprentissage automatique, en représentation de données et en évaluation expérimentale ; une appétence pour les approches multimodales et les LLM est attendue.
- Compétences techniques : bonne maîtrise de Python ; expérience appréciée avec les bibliothèques de deep learning, le traitement de documents et les bases de données graphes.
- Méthodologie de recherche : capacité à concevoir des protocoles expérimentaux rigoureux, à analyser les résultats de manière critique et à documenter les travaux dans une logique de reproductibilité.
- Compétences transversales : autonomie, esprit d'initiative, capacité de travail collaboratif et communication scientifique en français et en anglais (rédaction, présentations, échanges internationaux).

7 Ouverture internationale

Le projet s'inscrit dans une dynamique internationale, au croisement de la fouille de littérature scientifique, des modèles multimodaux et des graphes de connaissances. Les verrous scientifiques abordés (alignement inter-modalités, raisonnement traçable, évaluation de systèmes GraphRAG) concernent une communauté de recherche large au niveau international. Les résultats visés ont vocation à être diffusés dans des conférences et revues internationales de premier plan en apprentissage, modélisation et interrogation de données.

Des collaborations académiques seront recherchées avec des équipes travaillant sur l'analyse de documents scientifiques, l'interrogation de graphes et l'évaluation de LLM dans des contextes scientifiques. Ces collaborations pourront prendre la forme de co-encadrements ponctuels, de stages de recherche, de co-publications et de partages de jeux de données ou de bancs d'essais. Le projet favorisera également des interactions avec des infrastructures et initiatives de science ouverte (bases bibliométriques, plateformes de dépôts de code et de données), afin de renforcer l'interopérabilité des ressources, la diffusion des artefacts et l'impact international des contributions.

8 Conditions matérielles de réalisation du projet et conditions de sécurité spécifiques

Conditions matérielles : Le projet s'appuiera sur des stations de tavail équipées de GPU et le cluster Convergence du LIP6. L'environnement logiciel reposera sur des outils open source de traitement docu-

mentaire, d'apprentissage et de gestion de graphes, avec une gestion de versions systématique du code et des jeux de configurations expérimentales.

Conditions de sécurité spécifiques : Les travaux respecteront les règles institutionnelles en matière de sécurité des systèmes d'information et de protection des données : contrôle des accès aux serveurs et aux dépôts, sauvegardes régulières, chiffrement des données sensibles et traçabilité des manipulations expérimentales. Une vigilance particulière sera portée au respect des licences des jeux de données, aux contraintes de diffusion des artefacts (modèles, code, corpus) et à la conformité avec les principes d'éthique et de science ouverte applicables au projet.

9 Calendrier prévisionnel (3 ans)

- **Année 1 : Sélection des modèles et exploration des méthodes de fusion**
  - Revue de littérature sur l'alignement multimodal et le GraphRAG.
  - Évaluation, benchmark et prise en main des encodeurs SOTA pour les tables et les figures.
  - Définition formelle du problème d'alignement et caractérisation des topologies de graphes cibles.
- **Année 2 : Évaluation comparative et espace sémantique unifié**
  - Implémentation de plusieurs stratégies de fusion multimodale (ex. : approches contrastives, réseaux de projection, mécanismes d'attention).
  - Benchmark croisé de ces méthodes d'alignement en fonction des différents types de graphes à construire (graphes de connaissances, graphes de structure).
  - Validation sur des tâches de recherche d'information inter-modalités.
- **Année 3 : GraphRAG hétérogène et validation globale**
  - Construction du graphe documentaire multimodal et implémentation du pipeline GraphRAG.
  - Évaluation globale de l'écosystème sur des benchmarks (ex. : *Multimodal ArXiv* https://mm-arxiv.github.io/).
  - Rédaction du manuscrit et soutenance.

Références

[1] Yuhe Bai, Camélia Constantin, and Hubert Naacke. Leiden-fusion partitioning method for effective distributed training of graph embeddings. In *Machine Learning and Knowledge Discovery in Databases. Research Track - European Conference, ECML PKDD*, volume 14947, pages 366–382, 2024.

[2] Yuhe Bai, Modou Gueye, and Hubert Naacke. Selective multi-hop type-aware enhancement for context-limited knowledge graph entity typing. In *IEEE International Conference on Big Data*, pages 1–10, 2025.

[3] Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. Nougat : Neural optical understanding for academic documents. *arXiv preprint arXiv :2308.13418*, 2023.

[4] Allaa Boutaleb, Alaa Almutawa, Bernd Amann, Rafael Angarita, and Hubert Naacke. Hearts : Hypergraph-based related table search. In *ELLIS workshop on Representation Learning and Generative Models for Structured Data (RLGMSD)*, page 3, Amsterdam, NL, 2025. Poster.

[5] Allaa Boutaleb, Bernd Amann, Hubert Naacke, and Rafael Angarita. Something's Fishy In The Data Lake : A Critical Re-evaluation of Table Union Search Benchmarks. In *4th Table Representation Learning Workshop @ ACL 2025*, Vienna, Austria, 2025.

[6] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global : A graph rag approach to query-focused summarization. *arXiv preprint arXiv :2404.16130*, 2024.

[7] Negar Foroutan, Angelika Romanou, Matin Ansaripour, Julian Martin Eisenschlos, Karl Aberer, and Rémi Lebret. WikiMixQA : A multimodal benchmark for question answering over tables and charts. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics : ACL 2025*, pages 24941–24958, Vienna, Austria, July 2025. Association for Computational Linguistics.

[8] Santo Fortunato, Carl T Bergstrom, Katy Börner, James A Evans, Dirk Helbing, Staša Milojević, Alexander M Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, et al. Science of science. *Science*, 359(6379) :eaao0185, 2018.

[9] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models : A survey. *arXiv preprint arXiv :2312.10997*, 2023.

[10] Zirui Guo, Xubin Ren, Lingrui Xu, Jiahao Zhang, and Chao Huang. RAG-Anything : All-in-One RAG Framework. *arXiv preprint arXiv :2510.12323*, 2025.

[11] Feng Jiang, Kuang Wang, and Haizhou Li. Bridging research and readers : A multi-modal automated academic papers interpretation system. *arXiv preprint arXiv :2401.09150*, 2024.

[12] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9459–9474, 2020.

[13] Ke Li, Hubert Naacke, and Bernd Amann. An analytic graph data model and query language for exploring the evolution of science. *Big Data Research*, 26 :100247, 2021.

[14] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv : A dataset for improving scientific comprehension of large vision-language models. In *62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

[15] Zirui Li, Siwei Wu, Xingyu Wang, Yi Zhou, Yizhi Li, and Chenghua Lin. Docmmir : A framework for document multi-modal information retrieval. *arXiv preprint 2505.19312*, 2025.

[16] Fangyu Liu, Julian Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. Deplot : One-shot visual language reasoning by plot-to-table translation. In *Findings of the Association for Computational Linguistics : ACL 2023*, pages 10381–10399, 2023.

[17] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA : A benchmark for question answering about charts with visual and logical reasoning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics : ACL 2022*, pages 2263–2279, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[18] Jason Priem, Heather Piwowar, and Richard Orr. Openalex : A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv :2205.01833*, 2022.

[19] Ruyi Qi, Zhou Liu, and Wentao Zhang. DataCross : A Unified Benchmark and Agent Framework for Cross-Modal Heterogeneous Data Analysis. https ://arxiv.org/abs/2601.21403v1, January 2026.

[20] Hamed Rahimi, Hubert Naacke, Camélia Constantin, and Bernd Amann. ANTM : an aligned neural topic model for exploring evolving topics. In *Journées Bases de Données Avancées (BDA)*, page 11, Montpellier, France, 2023.

[21] Hamed Rahimi, Hubert Naacke, Camélia Constantin, and Bernd Amann. ATEM : A topic evolution model for the detection of emerging topics in scientific archives. In *12th International Conference on Complex Networks and their Applications*, page 10, Menton, France, 2023.

[22] Vicky Zayats, Kristina Toutanova, and Mari Ostendorf. Representations for question answering from documents with tables and text. In *16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume*, 2021.