PhD Topic

FUSION-KG: Framework for Unified multimodal Semantic extractION for Knowledge Graphs construction

Keywords: Vision-Language Models (VLMs); Ontology-Guided Multimodal Information Extraction; Knowledge Graphs; Hydro-ecological restoration.

This PhD research proposes the development of FUSION-KG, a unified framework for multimodal semantic extraction and structured knowledge graph construction from heterogeneous documentation sources. Building upon prior research on the extraction and structuring of restoration knowledge from textual corpus, particularly focused on restoration reports related to the Rhine River, the project aims to significantly extend this work by integrating visual and textual modalities within a coherent semantic modeling pipeline.

Research Context and Motivation

Environmental restoration projects generate large volumes of heterogeneous documentation, including technical reports, project plans, cartographic materials, engineering drawings, and photographic records [2]. These materials contain valuable but fragmented knowledge describing intervention strategies, environmental contexts, technical constraints, and outcomes.

Within the TETRA project[1], previous research efforts primarily concentrated on text-based knowledge extraction using Large Language Models (LLMs)[3], enabling the structuring of restoration knowledge from technical and narrative reports. While this approach demonstrated the potential of large language models for semantic modeling and ontology enrichment, it remained largely confined to textual sources. However, restoration documentation increasingly includes rich visual materials, such as maps, technical drawings, aerial imagery, and photographic records that contain complementary and sometimes critical information not explicitly described in text. This PhD builds upon the foundations established in TETRA by extending the extraction paradigm toward a unified multimodal framework. The central hypothesis is that integrating textual and visual understanding through advanced Vision-Language Models (VLMs) can substantially improve the completeness, semantic consistency, and interpretability of structured environmental knowledge graphs.

Recent advances in Large Language Models (LLMs) and Vision-Language Models (VLMs) have enabled effective extraction of entities and relations from textual and visual data[1, 12, 7, 6]. Yet, these models remain sensitive to noisy or unconventional inputs, particularly in specialized domains where textual formats are atypical and underrepresented in training corpus. Multimodal approaches that integrate visual and textual evidence offer greater semantic completeness and consistency, while the incorporation of structured external knowledge, such as ontologies, provides a mechanism to guide extraction, reduce ambiguity, and enforce conceptual coherence [11, 8, 9]. Despite these promising developments, methods for jointly aligning heterogeneous modalities and leveraging external knowledge for robust, domain-specific knowledge graph construction remain largely unexplored. The FUSION-KG framework aims to address this gap by unifying multimodal semantic extraction with ontology-driven knowledge graph construction, enabling more accurate, interpretable, and traceable representations of complex environmental documentation.

Methodology and scientific objectives

The FUSION-KG PhD aims to design a unified multimodal semantic extraction framework capable of transforming heterogeneous environmental documentation into structured, interpretable, and queryable knowledge graphs[4]. The ambition is not only to extract information from text and images, but to develop a coherent framework in which multimodal understanding and structured external knowledge jointly contribute to reliable and semantically consistent knowledge graph construction.

The work involves the systematic modeling and characterization of heterogeneous documentary sources, including technical reports, maps, engineering drawings, aerial and satellite imagery, and photographic records of restoration interventions. These materials provide complementary yet often fragmented accounts of intervention types, spatial configurations, temporal phases, environmental parameters, constraints, and outcomes. A major challenge lies in ensuring that information extracted from visual and

[1] `https://anr.fr/Project-ANR-22-FAI2-0006`

textual modalities is semantically aligned and represented within a shared conceptual framework. In this context, Research Question 1 (**RQ1**) examines how cross-modal alignment can be achieved so that visual elements and textual descriptions are consistently interconnected during knowledge graph construction. Addressing this question requires the design of alignment strategies capable of grounding extracted entities and relations simultaneously in visual evidence, textual references, and the underlying conceptual structure (ontology), thereby preserving semantic coherence across modalities.

At the methodological core of FUSION-KG is the development of multimodal semantic extraction pipelines based on Large Language Models (LLMs) and Vision-Language Models (VLMs). However, it is important to recognize inherent limitations of current LLMs. These models are sensitive to the noisy and underspecified nature of prompts, and their lack of deep contextual understanding can lead to linguistic ambiguities[5, 10]. Because LLMs primarily rely on statistical patterns learned from large corpus rather than explicit reasoning mechanisms, they may generate plausible but semantically imprecise or inconsistent outputs. This vulnerability becomes particularly critical in domain-specific knowledge extraction, where precision, disambiguation, and conceptual consistency are essential. Moreover, this issue is especially critical in domains characterized by unconventional or domain-specific textual formats, for which LLMs lack prior exposure in their training corpus.

For this reason, the project explicitly emphasizes the integration of external structured knowledge, particularly domain ontologies, into the extraction and graph construction process. Ontologies will provide formal definitions of domain concepts, properties, and relations, guiding entity recognition, relation extraction, normalization, and semantic validation. By introducing structured conceptual constraints, the framework aims to reduce ambiguity, mitigate prompt sensitivity, and enforce domain coherence throughout the extraction pipeline. This leads to Research Question 2 (**RQ2**): how can external structured knowledge, such as ontologies, be effectively integrated into multimodal LLM and VLM pipelines to guide and constrain semantic information extraction? This question explores mechanisms such as ontology-informed prompting, schema-guided extraction templates, vocabulary alignment, and structured validation procedures that embed domain knowledge directly into the modeling process.

Beyond methodological integration, the project also investigates the measurable impact of ontology-guided multimodal extraction on the quality of the resulting knowledge graphs. Thus, Research Question 3 (**RQ3**) examines to what extent ontology-guided multimodal extraction improves the completeness, reliability, and explainability of constructed knowledge graphs compared to text-only or unconstrained multimodal approaches. This includes assessing whether structured knowledge integration reduces semantic inconsistencies, enhances entity disambiguation, limits the propagation of ambiguous interpretations, and improves traceability between extracted triples and their documentary sources.

The extracted information will be formalized as RDF-based knowledge graphs compliant with the domain ontology. Automated pipelines will transform multimodal extraction outputs into semantically validated triples, enabling querying, reasoning, comparison across restoration cases, and longitudinal analysis. Particular attention will be given to traceability, ensuring that each graph assertion can be linked back to its original textual or visual evidence.

Evaluation will combine technical and functional dimensions. On the technical side, the research will assess entity and relation extraction performance, cross-modal alignment accuracy, ontology compliance and robustness to prompt variation. On the functional side, it will evaluate the usefulness of the resulting knowledge graphs for domain experts, particularly their ability to answer competency questions, support comparative evaluation of restoration strategies, and facilitate evidence-based decision-making.

Through this integrated framework, FUSION-KG seeks to demonstrate that the integration of external structured knowledge into multimodal extraction pipelines can substantially advance the robustness, interpretability, and practical value of knowledge graph construction from complex environmental documentation.

Work Environment

This PhD will be carried out mainly in the SDC team of the ICube laboratory (CNRS UMR 7357). The PhD will begin in september 2026 for an estimated period of 36 months. The candidate will be welcomed into the ICube laboratory and supervised by Florence Le Ber and Franco Giustozzi.

Candidate

Specific knowledge: Knowledge on data science methods, knowledge representation and reasoning, knowledge graphs. Languages: Python, java, owl/sparql. Ability to work with experts who are not computer scientists. Interest in the application domain would be appreciated.

Education: Student about to graduate a Master or Engineer (Bac + 5) with a specialization in Computer Science.

How to apply

The interested candidates must send an email to Florence Le Ber (`florence.leber@engees.unistra.fr`) and Franco Giustozzi (`franco.giustozzi@insa-strasbourg.fr`) with the following documents:

- A CV,
- A cover letter (max. 1 page), including the applicant's motivation for applying and a brief explanation of their academic background,
- A transcript of the available grades for the current year and the past year.

References

[1] Florian Bordes and et.al. An introduction to vision-language modeling. *CoRR*, abs/2405.17247, 2024.

[2] Valentin Chardon, Cassandra Euzen, and Laurent Schmitt. Functional river restoration as a lever for adapting to climate change from an interdisciplinary emblematic showcase on the upper rhine. *Journal of Environmental Management*, 393:127151, 2025.

[3] Fethi Ghazouani, Franco Giustozzi, and Florence Le Ber. Llm-driven case-base populating for structuring and integrating restoration experiences. In Isabelle Bichindaritz and Beatriz López, editors, *Case-Based Reasoning Research and Development*, pages 67–80, Cham, 2025. Springer Nature Switzerland.

[4] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, et al. Knowledge graphs. *ACM Computing Surveys (Csur)*, 54(4):1–37, 2021.

[5] Aryan Keluskar, Amrita Bhattacharjee, and Huan Liu. Do llms understand ambiguity in text? a case study in open-world question answering, 2024.

[6] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[7] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[8] Zhangcheng Qiang, Kerry Taylor, Weiqing Wang, and Jing Jiang. Oaei-llm: a benchmark dataset for understanding large language model hallucinations in ontology matching. *arXiv preprint arXiv:2409.14038*, 2024.

[9] Xiaoyang Wei, Zografoula Vagena, Camille Kurtz, and Florence Cloppet. Integrating expert knowledge with vision-language model for medical image retrieval. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–4, 2024.

[10] Shuguang Yang, Feipeng Chen, Yiming Yang, and Zude Zhu. A study on semantic understanding of large language models from the perspective of ambiguity resolution. In *Proceedings of the 2023 International Joint Conference on Robotics and Artificial Intelligence*, JCRAI '23, page 165–170, New York, NY, USA, 2024. Association for Computing Machinery.

[11] Weiqi Ye, Qiang Zhang, Xian Zhou, Wenpeng Hu, Changhai Tian, and Jiajun Cheng. Correcting factual errors in llms via inference paths based on knowledge graph. In *2024 International Conference on Computational Linguistics and Natural Language Processing (CLNLP)*, pages 12–16. IEEE, 2024.

[12] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.