OntoSepsisKG : Omics-based Knowledge Graph for Sepsis

Laura FORERO CAMACHO*,**, Nacéra SEGHOUANI*, Gregoire TOURNOIS*, Farida ZEHRAOUI**

*LISN, Université Paris-Saclay, Paris, France
{laura-isabella.forero-camacho, nacera.seghouani}@centralesupelec.fr
gregoire.tournois@universite-paris-saclay.fr
**IBISC, Université d'Evry, Paris-Saclay, Paris, France
farida.zehraoui@univ-evry.fr

Résumé. Cet article présente une méthodologie appelée OntoSepsisKG pour construire un graphe de connaissances basé sur l'ontologie omique pour la septicémie. OntoSepsisKG comprend deux parties principales : (i) l'intégration d'ontologies; et (ii) la création d'instances à partir de données patients et profiles d'expression génique. La tâche principale de cette partie consiste à identifier les interactions entre les gènes et les protéines. L'analyse a révélé des différences significatives dans l'expression des gènes entre les groupes de patients atteints de sepsis (survivants et non) et le groupe sain, en particulier dans les gènes associés au dysfonctionnement mitochondrial et à la dysrégulation immunitaire.

1 Introduction

Sepsis is a life-threatening medical condition defined as life-threatening organ dysfunction caused by a dysregulated host response to infection Singer et al. (2016). The body's reaction can damage tissues and organs, leading to shock, multiple organ failure, and high mortality. Evidence suggests that Sepsis is a complex condition not only due to its ability to be caused by various infectious agents and immune responses, as outlined by its pathology, but also due to significant alterations in coagulation, immunosuppression, and organ dysfunction Gyawali et al. (2019). These factors underscore the importance of early identification and treatment, emphasizing the need to recognize at-risk patients for the design of effective, personalized therapies that target specific mechanisms. Recent research has leveraged omics data to identify key biological pathways involved in the pathogenesis of Sepsis Gyawali et al. (2019). omics is an integrative approach in biological research that combines data from multiple "omics" fields, such as genomics (study of DNA), transcriptomics (study of RNA), proteomics (study of proteins), metabolomics (study of metabolites), and others. Several studies, including those by Xu et al. (2021) and Zhang et al. (2020), emphasize the potential of omics data integration and ontology-based approaches in the analysis of complex diseases. In the case of Sepsis, omics play a critical role in patient profiling. The use of gene expression data (Transcriptomics) collected in blood-based tests offers more accessible and cost-effective alternatives that may improve the identification process of Sepsis. However, challenges remain in harmonizing

datasets provided by different sources due to variations in data semantics, and experimental protocols.

The ontology is a formal representation of specific domain knowledge. It enables the integration of heterogeneous data sources and supporting reasoning and extensible frameworks. For example, Gene Ontology (GO) Ashburner et al. (2000) provides structured knowledge useful for biological analysis, as in Sepsis profiling studies (Zhang et al. (2020)). Recent studies, such as Lamy (2017) define a framework for biomedical ontologies, and show the relevance of KGs in the identification of therapeutic targets and repurposing drugs for complex diseases. Similarly, Li et al. (2021) leverages the Protein-Protein Interaction Ontology (PPIO) for integrating omics data.

As part of IHU Prometheus [1] project, this work aims to define a methodology called OntoSepsisKG to build a knowledge graph based on omics ontology for Sepsis. The primary data source is transcriptomics, which provides data on gene expression and protein activity. The methodology was structured into two main parts. The first focused on integrating external ontologies, such as Gene Ontology (GO). The second involved creating instances using the datasets, which included patient, sample data, and gene expression profiles. This step encompassed identifying Sepsis-related genes and enriching the KG by incorporating external databases and ontology information. This enables the identification of key pathways involved in Sepsis, including immune system regulation and signal transduction. The OntoSepsisKG framework illustrates the relevance of integrating omics data with curated domain knowledge to contextualize gene expression changes and provide deeper insights into the molecular mechanisms of Sepsis.

This paper is structured as follows : Section 2 describes the methodology used to construct the knowledge graph, while Section 3 presents the results of the preliminary analysis. Finally, section 4 concludes the paper and discusses its limitations.

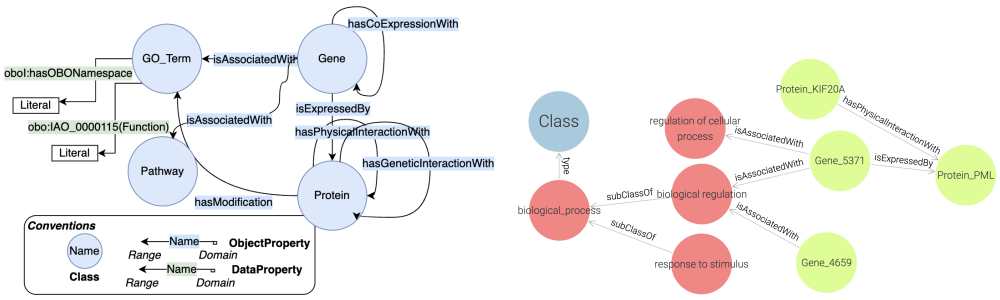
2 OntoSepsisKG Methodology

2.1 Main Concepts Related to OntoSepsisKG

Gene Ontology (GO) Gene Ontology provides a structured framework for annotating genes, including three main categories : Molecular Function (MF), Biological Process (BP), and Cellular Component (CC). These categories ensure clear and consistent descriptions of gene functions, making it easier to compare and standardize data across databases. As illustrated in Figure 1b, GO terms are used to enrich the biological context of the genes analyzed in this paper. For example, gene 5371 is associated with the biological regulation process within GO framework. This association will aid in interpreting the gene's potential role in Sepsis.

Integration of GO in OntoSepsisKG To integrate information from GO, four classes are defined : *Gene*, *Protein*, and *Pathway*, as shown in Figure 1a. *Gene* represents a gene identified by an ID. To create relationships between genes with similar expression patterns, we defined the object property *hasCoExpressionWith*. Similarly, *isExpressedBy* object property defines a relation between *Gene* and *Protein*. *Protein* can interact with other proteins through two types of relationships : *hasGeneticInteractionWith* and *hasPhysicalInteractionWith*. A gene can be associated with a GO term, corresponding to one of the classes defined in GO.

1. https://www.fhu-Sepsis.uvsq.fr/ihu-prometheus-1



(a) Integration of GO in OntoSepsisKG (b) Instances Gene Expression on OntoSepsisKG

FIG. 1 – *Main Concepts Related to OntoSepsisKG*

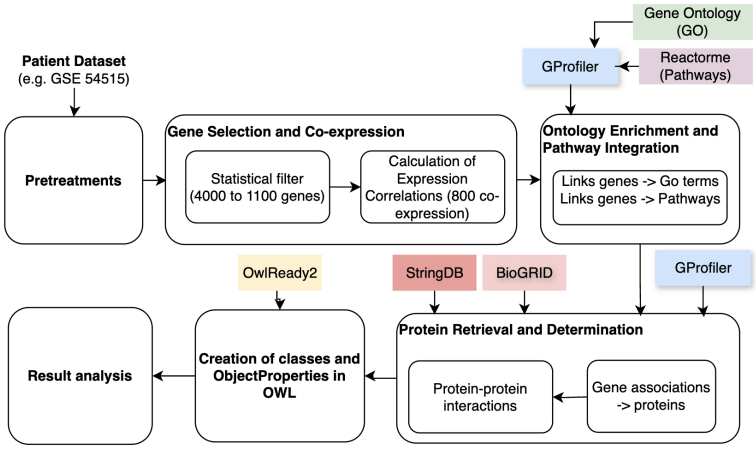


FIG. 2 – *Process of creation of Instances for each dataset*

2.2 Creation of Knowledge Graph

The construction of the Sepsis knowledge graph follows a systematic methodology that involves several key steps, as depicted in Figure 2. Each step is designed to ensure that the resulting KG accurately reflects the biological processes and interactions relevant to Sepsis.

2.2.1 Patient Datasets and Pre-treatments

The datasets used for creating the KG include GSE54514, GSE64456, and GSE57065, all available on the NCBI GEO platform. GSE54514 (Parnell et al. (2013)) includes transcriptomic profiles from whole blood samples of Sepsis patients, divided into survivors (n=26), nonsurvivors (n=9), and healthy controls (n=18), with a total of 53 patients. GSE64456 (Mahajan et al. (2016)) contains 298 samples, classified into bacterial infections (n=89) and non-bacterial infections (n=190). This dataset is particularly significant as bacterial infections represent a ma-

jor cause of Sepsis. GSE57065 (Tabone et al. (2018)) investigates the genomic response of 28 ICU patients at the onset of septic shock, a critical condition in which Sepsis triggers severe circulatory, cellular, and metabolic abnormalities, leading to life-threatening organ dysfunction. Blood samples were collected at three time points, focusing on early genomic changes and their potential impact on patient outcomes.

For the creation of the KG, classes *Dataset*, *Patient* and *Patient Sample* were defined. A patient can be associated with multiple samples, each containing details like collection type, date, and tissue site. Patient samples are linked to genes through the *hasExpressionValue* object property, enabling the determination of gene expression values for each sample and gene.

Extraction and Importance of Patient Metadata The extraction of patient metadata from the datasets is essential for accurately characterizing and analyzing the data. Data such as tissue type, disease state, and patient demographics (e.g., age and gender) are examples of metadata. The contextualization of gene expression data within biological and clinical aspects is aided by this information. We enhance the KG and enable searches that integrate patient-specific information with transcriptomics data by connecting these metadata to patient instances.

Handling Inconsistent Data (Pre-treatments) Managing discrepancies in data formats and structures is a crucial difficulty when merging several datasets. The formats used for gene identifiers, probe mappings, and metadata may differ throughout GEO datasets. Furthermore, datasets may use various platforms (e.g., Illumina, Affymetrix) or technologies (e.g., microarrays or RNA-seq), which might result in variations in data structure and annotation. We resolved these discrepancies in our work by applying unique parsing strategies for every dataset. For example, we created techniques to standardize metadata fields across datasets and deal with situations where probes map to several genes (for instance, by utilizing the explode function).

2.2.2 Gene Selection and Co-Expression

Filtering Genes Given the high dimension of a single dataset in terms of genes, we decided to apply an initial filter to identify genes that are statistically significant in the context of Sepsis. The GEOparse tool [2] was used to process the gene expression data. The probes were mapped to Entrez gene IDs [3] using the GPL annotations [4] provided by the probe constructor. Differentially expressed genes (DEGs) were defined as those with an adjusted p-value of less than 0.01 in a statistical test (T-test), following a method similar to the one used in the study Kim et al. (2021). This filtering phase reduces the dimensions of the data and ensures that the dataset contains only the most relevant genes for the Sepsis analysis.

Co-expression-Based Interaction To establish relationships between two genes, a Spearman correlation matrix was calculated for the differentially expressed genes (DEGs) to identify coexpression patterns. The Spearman correlation helps identify genes that exhibit similar expression patterns. Genes with a correlation coefficient greater than 0.8 were considered co-expressed. The coexpressed genes were directly linked using the *hasCoExpressionWith* property in the ontology.

2. Python library to access Gene Expression Omnibus Database (GEO).
3. Entrez gene IDs are unique identifiers assigned by the NCBI to genes for standardized referencing.
4. GPL annotations refer to platform-specific metadata files provided by microarray manufacturers, which include probe information and corresponding gene identifiers.

Ontology Enrichment and Pathway Integration

Linking with GO GO annotations for the selected genes were recovered using GProfiler. In this study, the three main GO categories were applied to the differentially expressed genes (DEGs). For example, GO terms related to "immune system process" (Term in BP), "receptor activity" (Term in MF), and "plasma membrane" (Term in CC) help identify genes involved in immune response. This annotation linking ensures consistent categorization of genes.
Pathway Retrieval Reactome [5] is an open-source resource that provides detailed information on biological pathways, combining data from experimental studies and scientific literature. GO annotations provide context for understanding biological processes, while Reactome delivers more detailed information about specific pathways and their roles in gene expression. In this phase, the focus was on linking genes to pathways, retrieved from Reactome, using the *isAssociatedWithPathway* relationship. Future steps will expand the use of Reactome to include more complex details, such as biochemical reactions and protein interactions, to provide a more complete understanding of the biological processes involved.

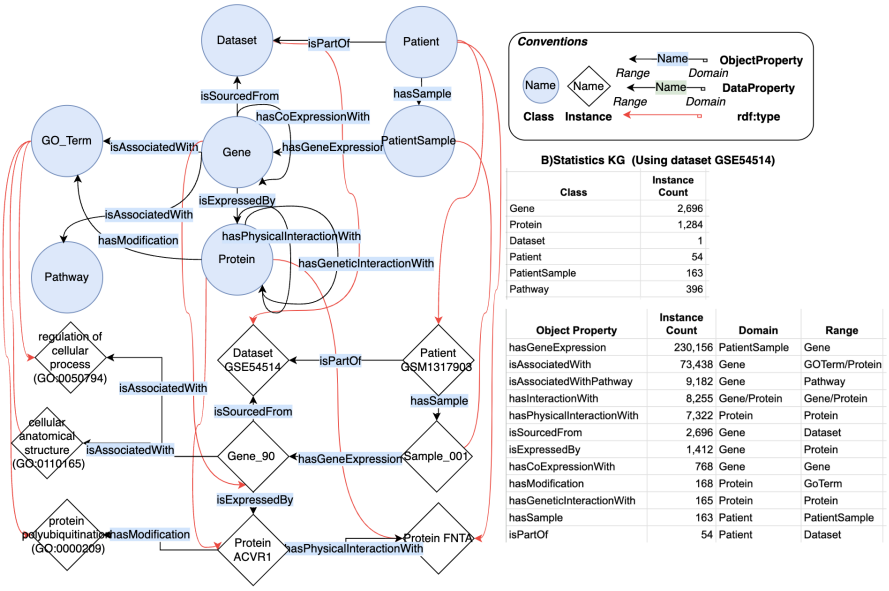
2.2.3 Protein Retrieval and Determination

Following the creation of genes, protein information was integrated using the STRING [6] and BioGRID [7] databases, which provide data on protein-protein interactions (PPIs) and modifications. STRING was queried with a list of protein names to retrieve interactions and their confidence scores. Moreover, BioGRID data were filtered to include specific modifications and details about the interactions. The inclusion of *hasModification* not only brings additional data about the protein but also enables relationships with GO to encompass different terms. Proteins associated with the selected genes were instantiated, and their interactions were represented using the *hasPhysicalInteractionWith* and *hasGeneticInteractionWith* object properties.

3 Preliminary Analysis

An OWL ontology was developed to model the classes of *Gene*, *Protein*, *Dataset*, *Patient*, *PatientSample*, and *Pathway*, with object properties defining their relationships. The resulting KG and its statistics are summarized in Figure 3. The pipeline is designed to be extendable, enabling its application to future datasets. The purpose of this analysis is to validate the quality of the KG. This included examining gene expression differences to ensure the proper application of filtering strategies. Furthermore, we explored the biological terms and pathways associated with genes, confirming the identification of mechanisms underlying Sepsis. SPARQL queries were utilized to assess the methodology applied in constructing the KG.
Dataset Description The analysis focused on the GSE54514 dataset to validate the quality of the KG and its information, as this dataset provided a more diverse distribution of patients and detailed information related to Sepsis. The dataset comprises data from 54 participants divided into three groups : Sepsis survivors (48.1%), healthy controls (33.3%), and Sepsis nonsurvivors (18.5%). Participant ages ranged from 18 to 86 years, with most individuals being middle-aged

5. `https://reactome.org/`
6. `https://string-db.org/`
7. `https://thebiogrid.org/`



B)Statistics KG (Using dataset GSE54514)

Class	Instance Count
Gene	2,696
Protein	1,284
Dataset	1
Patient	54
PatientSample	163
Pathway	396

Object Property	Instance Count	Domain	Range
hasGeneExpression	230,156	PatientSample	Gene
isAssociatedWith	73,438	Gene	GOTerm/Protein
isAssociatedWithPathway	9,182	Gene	Pathway
hasInteractionWith	8,255	Gene/Protein	Gene/Protein
hasPhysicalInteractionWith	7,322	Protein	Protein
isSourcedFrom	2,696	Gene	Dataset
isExpressedBy	1,412	Gene	Protein
hasCoExpressionWith	768	Gene	Gene
hasModification	168	Protein	GoTerm
hasGeneticInteractionWith	165	Protein	Protein
hasSample	163	Patient	PatientSample
isPartOf	54	Patient	Dataset

FIG. 3 – *Overview of KG, Instance creation and statistics*

or older. Regarding infection sites, 40% of the Sepsis cases were linked to lung infections, 15% to bloodstream infections, 12.5% to urinary tract infections, and 25% classified as "Other".

Quantification of Gene Expression Differences Across Sepsis and Control Groups We validated the methodology used to construct the KG by employing SPARQL queries to analyze gene expression data from Sepsis groups (survivors and non-survivors) and the control group. As expected, we observe that the average gene expression levels in the Sepsis groups is higher than those in the control group for the selected genes. This confirms that the significance filter applied during gene selection is effectively implemented. Among the most notable differences, we find genes associated with "Regulation of actin polymerization or depolymerization," "Regulation of actin filament length," and "Obsolete respirasome." These findings suggest that processes such as cytoskeletal regulation and mitochondrial dysfunction are more pronounced in Sepsis. This analysis supports the hypothesis that Sepsis results in distinct gene expression changes when compared to healthy controls, consistent with the findings in similar studies, such as the one by Kim et al. (2021).

Gene Ontology Terms and Their Biological Relevance We analyzed the Gene Ontology (GO) terms linked to the most highly expressed genes in both the Sepsis and control groups. In the Sepsis and non-survivor groups, the most significant terms were "obsolete respirasome" and "molecular sequestering activity." The term "obsolete respirasome" refers to mitochondrial dysfunction, a key feature of Sepsis that contributes to organ damage. Targeting mitochondrial function has been suggested as a potential therapeutic approach to restore energy balance and reduce inflammation in Sepsis Li et al. (2022), and may also aid in early Sepsis detection. In contrast, the control group's top genes were associated with terms like "molecular sequestering activity," "focal adhesion," and "cell-substrate junction," which reflect biological processes less

disrupted in healthy individuals. These findings support the hypothesis that genes involved in mitochondrial dysfunction and actin regulation are more highly expressed in Sepsis compared to controls, indicating disease-related biological changes. This observation is consistent with previous studies, such as Kim et al. (2021), that highlight the role of mitochondrial dysfunction in Sepsis progression.

Enrichment Through Pathway Integration We hypothesized that adding Reactome pathways would improve the biological relevance and specificity of the KG's associations, particularly for Sepsis-related genes. Upon examining the data, we found that key pathways related to the Immune System, Signal Transduction, and the Innate Immune System were highly represented, all of which are critical for understanding the immune dysregulation that occurs in Sepsis. Additionally, pathways involved in the Metabolism of Proteins and Cellular Responses to Stress are important for understanding the cellular and tissue-level impacts of Sepsis. These findings highlight the relevance of including this information, which was not captured by GO annotations, especially for the identification of pathways involved in the progression and response to Sepsis.

4 Conclusions and Perspectives

In this study, we developed a methodology for creating a Knowledge Graph (KG) that integrates omics data to better understand Sepsis. The approach involved using transcriptomics datasets along with biological ontologies such as Gene Ontology (GO) and biological databases. The analysis showed clear differences in gene expression between Sepsis groups and healthy controls, confirming that Sepsis leads to distinct molecular changes. Specifically, genes associated with mitochondrial dysfunction, cytoskeletal regulation, and immune dysregulation were more highly expressed in Sepsis. Furthermore, the addition of biological databases and GO further improved the KG's biological relevance, helping to identify important pathways like "Signal Transduction" and "Immune System", which are key in Sepsis progression. The combination of omics data with existing biological knowledge allowed us to better understand the gene expression changes in Sepsis. The framework we created offers a useful tool for studying Sepsis, making it easier for researchers to integrate different datasets and gain a better understanding of the biological processes involved. While this study focused mainly on transcriptomics data, future work will aim to include multiple omics, such as proteomics and metabolomics, to further improve the KG and deepen understanding of Sepsis.

Références

Li, W. et al. (2022). Mechanism of mitophagy and its role in sepsis induced organ dysfunction : A review. *International Journal of Molecular Sciences 23*(3), 992, doi:10.3390/ijms23030992.

Li, M., Q. He, C. Yang, J. Ma, F. He, T. Chen, et Y. Zhu (2021). The protein-protein interaction ontology. *BMC Genomics 22 (Suppl 5)*, 544.

Kim, K., D. Jekarl, J. Yoo, S. Lee, M. Kim, et Y. Kim (2021). Immune gene expression networks in sepsis : A network biology approach. *PLoS One 16*(3), e0247669, doi:10.1371/journal.pone.0247669.

Xu, N., G. Hui, L. Xurui, Z. Qian, et L. Jianguo (2021). A five-genes based diagnostic signature for sepsis-induced ards. *Pathology and Oncology Research 27*.

Zhang, Z., C. Lin, P. Xu, L. Xing, Y. Hong, et P. Chen (2020). Gene correlation network analysis to identify regulatory factors in sepsis research. *Journal of Translational Medicine 18*(381).

Gyawali, B., K. Ramakrishna, et A. S. Dhamoon (2019). Sepsis : The evolution in definition, pathophysiology, and management. *SAGE Open Medicine 7*, 2050312119835043, doi:10.1177/2050312119835043.

Tabone, O., M. Mommert, C. Jourdan, E. Cerrato, et al. (2018). Endogenous retroviruses transcriptional modulation after severe infection, trauma and burn. *Frontiers in Immunology 9*, 3091, doi:10.3389/fimmu.2018.03091.

Lamy, J.-B. (2017). Owlready : Ontology-oriented programming in python with automatic classification and high level constructs for biomedical ontologies. *Artificial Intelligence in Medicine 80*, 11–28.

Singer, M., C. S. Deutschman, C. W. Seymour, et al. (2016). The third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA 315*, 801–810.

Mahajan, P., N. Kuppermann, A. Mejias, N. Suarez, D. Chaussabel, T. C. Casper, J. M. Dean, et O. Ramilo (2016). Association of rna biosignatures with bacterial infections in febrile infants aged 60 days or younger. *JAMA 316*(8), 846–857, doi:10.1001/jama.2016.9207.

Parnell, G. P., B. M. Tang, M. Nalos, N. J. Armstrong, et al. (2013). Identifying key regulatory genes in the whole blood of septic patients to monitor underlying immune dysfunctions. *Shock 40*(3), 166–174, doi:10.1097/SHK.0b013e31829965c2.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, et al. (2000). Gene ontology : tool for the unification of biology. *Nature Genetics 25*(1), 25–29, doi:10.1038/75556.

Summary

This paper presents a methodology called OntoSepsisKG to build a knowledge graph based on omics ontology for Sepsis. The primary data source is transcriptomics, which provides data on gene expression and protein activity. OntoSepsisKG KG includes two main parts: (i) Integration of external ontologies, such as Gene Ontology (GO). (ii) Creation of instances using datasets about different samples and gene expression profiles. The main task in this part is how to identify interactions between genes and proteins using these profiles. The analysis revealed significant gene expression differences between Sepsis patient groups (survivors and not survivors) and the healthy group, particularly in genes associated with mitochondrial dysfunction and immune dysregulation. These preliminary results show the relevance of integrating omics ontology with curated domain knowledge to contextualize gene expression changes and provide deeper insights into the molecular mechanisms of Sepsis.

Apport des LLMS pour peupler une ontologie de restauration des hydro-écosystèmes

Fethi Ghazouani*, Franco Giustozzi**
Florence Le Ber***

Université de Strasbourg, ENGEES, INSA, CNRS, ICube UMR 7357, F-67000 Strasbourg, France
*fghazouani@unistra.fr, **franco.giustozzi@insa-strasbourg.fr,
***florence.leber@engees.unistra.fr

Résumé. La restauration des écosystèmes hydrologiques est essentielle pour maintenir la qualité et la durabilité des ressources en eau. Les gestionnaires d'infrastructures hydrauliques, barrages ou centrales hydroélectriques, ont l'obligation légale de renaturaliser une partie de leur zone d'action afin de contribuer à ces efforts. Plusieurs expériences de restauration ont ainsi été menées, mais sans retour d'expérience systématique, ce qui limite la capacité des gestionnaires à planifier et à exécuter efficacement les nouveaux projets. Cet article présente une approche innovante pour structurer et exploiter ces données grâce au peuplement automatique d'une ontologie de domaine, assisté par des grands modèles de langage (LLMs) [1].

1 Introduction

La restauration des hydro-écosystèmes représente un enjeu majeur pour la gestion durable des ressources en eau, ce qui justifie les exigences légales qui régissent les opérations de restauration des barrages et des centrales hydroélectriques. Ces exigences imposent une renaturalisation des zones d'action, qui ne peut être efficacement réalisée sans une connaissance approfondie des projets de restauration passés. Cependant, malgré les efforts entrepris pour collecter des retours d'expérience sur ces opérations de restauration Schmitt et Beisel (2015), les données sont souvent peu exploitables. Ce constat révèle le besoin d'une approche structurée pour analyser et intégrer les expériences de restauration, ce qui a conduit à envisager un cadre de raisonnement à partir de cas (Aamodt et Plaza, 1994), où les expériences sont formalisées comme des cas réutilisables, et où les connaissances du domaine sont formalisées dans une ontologie. Développée pour représenter les concepts et les relations propres aux opérations de restauration, cette ontologie permettrait une analyse systématique des retours d'expérience et offrirait un référentiel structuré pour guider de futurs projets de restauration. L'instanciation des cas, ou peuplement de l'ontologie, s'appuie ici sur principalement sur des rapports d'opérations ou des fiches décrivant les projets. Dans cet article, nous abordons spécifiquement le peuplement automatique de l'ontologie à l'aide de grands modèles de langage (*Large Language Models*,

1. Ces travaux sont financés par le projet ANR-DLR Tetra (ANR-22-FAI2-0006-02).

ou LLMs) à partir des textes peu structurés. Ces modèles entraînés sur de vastes ensembles de données, sont capables de générer automatiquement des triplets pour enrichir l'ontologie.

La suite de article positionne notre approche vis-à-vis des travaux connexes (section 2) ; la section 3 présente l'ontologie et la stratégie de peuplement à partir des LLMs, tandis que la section 4 détaille les résultats obtenus. La dernière section est une conclusion.

2 Etat de l'art

Nous traitons ici des approches qui visent à extraire des informations à partir de documents textuels non structurés pour enrichir une ontologie de domaine. Lubani et al. (2019) classent les techniques de peuplement d'ontologies selon deux grandes approches : le peuplement d'ontologie d'un point de vue sémantique et le peuplement d'ontologie d'un point de vue "Intelligence Artificielle (IA) générative" (Sahbi et al., 2024). Pour la première catégorie, l'accent est mis sur l'identification et la structuration des concepts et relations de manière à refléter précisément le domaine de connaissances. Ces approches, qui reposent généralement sur des algorithmes d'apprentissage ou des règles nécessitent une supervision humaine importante ou le développement de règles spécifiques pour chaque application.

Le peuplement d'ontologie d'un point de vue IA générative exploite les capacités des modèles de langage pour générer automatiquement des instances et des assertions de propriétés, facilitant ainsi une intégration dynamique et évolutive des connaissances. Les innovations en IA ont permis l'émergence de grands modèles de langage, tels que GPT d'OpenAI (OpenAI et al., 2023), LlaMA de Meta (Touvron et al., 2023), Qwen d'Alibaba Cloud (Bai et al., 2023), etc. Ces modèles, entraînés sur de vastes quantités de données non étiquetées grâce à un pré-entraînement auto-supervisé à grande échelle, s'adaptent à une grande variété de tâches. En particulier, es LLMs ont montré des performances avancées dans les tâches d'extraction de connaissances, comme le souligne (Li et al., 2023).

L'apprentissage par contexte (In-Context Learning, (ICL) (Brown, 2020)) où le modèle apprend à reproduire des tâches en à partir d'exemples est particulièrement bénéfique pour les LLM, qui, grâce à leur vaste nombre de paramètres, peuvent exécuter des tâches en répondant à des entrées soigneusement conçues, telles que des prompts textuels adaptés. L'ingénierie des prompts s'est alors développée pour aider à concevoir des prompts adaptés. Par exemple, Polat et al. (2024) ont utilisé diverses techniques d'ingénierie des prompts pour extraire des connaissances à partir de textes. Kouhoue et al. (2024) exploite *Claude* pour peupler une ontologie à partir d'annonces immobilières. Dans (Sahbi et al., 2024), les auteurs évaluent les performances des modèles *ChatGPT* et *TextCortex* pour peupler une ontologie de maintenance des bâtiments à partir de données CSV semi-structurées.

Ces différents travaux servent de guide à l'approche proposée pour peupler l'ontologie sur la restautaion des hydro-écosystèmes.

3 Méthodologie

TetraOnto, une ontologie sur la restauration des hydro-écosystèmes. L'ontologie dénommée TetraOnto, a été construite à partir de données provenant de diverses sources, telles que des retours d'expérience de restauration, des rapports d'opérations et des fiches d'interviews. Ces

documents contiennent des informations essentielles sur les opérations de restaurations réalisées, les zones géographiques concernées, les espèces impactées, les infrastructures présentes (e.g., barrages) et les résultats des interventions. Ce travail a été mené en étroite collaboration avec des experts du domaine afin d'assurer une représentation précise et cohérente des connaissances spécifiques au contexte de la restauration des hydro-écosystèmes. L'ontologie construite (voir figure 1) intègre des concepts tels que les types de mesures de restauration (*restorationMeasure*) réalisées, comme la reconnexion (reconnection), la création de bassins (*poolCreation*) ou la restauration de berges ; les types de cours d'eau (*waterBody*) tels que les rivières (*river*) ou les bras d'eau (*pool*) ; les structures (*structure*) impliquées, tels que les barrages (*dam*) et les seuils (*weir*) ; les zones géographiques (*geographicZone*) ; les espèces migratrices (*migratorySpecie*) ; et le maître d'oeuvre d'une opération de restauration (*mainContractor*). Elle inclut des relations entre concepts, telles que '*realizedIn*', reliant une mesure de restauration à une zone géographique, '*associatedTo*', qui associe une mesure de restauration à une structure, ou encore '*isLocatedOn*', qui localise une structure dans un cours d'eau. Enfin, des propriétés de données comme '*hasCost*' (coût d'une mesure de restauration), '*startsAt*' ou '*hasHeight*' (hauteur d'une structure) permettent de documenter des informations quantitatives et temporelles sur les opérations de restauration.

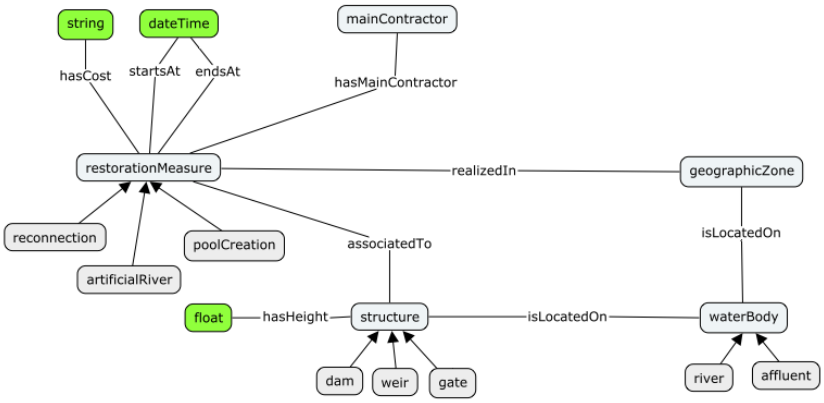


FIG. 1 – *Un extrait de l'ontologie de restauration d'hydrosystèmes*

Peuplement à l'aide de LLMs. Nous considérons le peuplement de l'ontologie comme une tâche d'extraction générative de connaissances, visant à identifier et structurer les entités ainsi que leurs relations sous forme de triplets. Un LLM est utilisé comme extracteur des triplets cibles à partir des textes, en s'appuyant sur l'ontologie de domaine pour guider l'identification et la structuration des entités et des relations pertinentes. L'entrée du modèle est un prompt contenant la description de la tâche de peuplement de l'ontologie (DT), l'ontologie (O) et le texte d'entrée (TE), avec un ensemble facultatif d'exemples ([EX]) illustrant le processus d'extraction/peuplement sur différents textes d'entrée. Le processus de peuplement est formulé

comme suit :

$$\text{LLM}(\text{Prompt}(DT, [\text{EX}], O, IT)) = [(h, r, t)_0, (h, r, t)_1, \ldots, (h, r, t)_i] \quad (1)$$

où $(h, r, t)_i$ est un triplet extractible du texte d'entrée et i le nombre total de triplets prédits.

Une étape d'ingénierie des prompts est mise en place pour identifier le prompt le plus adapté à la tâche visée. Nous utilisons pour cela deux approches principales, adaptées à notre contexte comme suit : *1)* Prompt Direct se réfère à un prompt qui ne comprend que la description de la tâche et l'entrée sur laquelle travailler ; *2)* l'apprentissage par contexte (ICL) qui ajoute un exemple de solution à la tâche visée sur une donnée d'entrée différente. Ces deux approches sont expérimentées avec plusieurs LLMs. Les résultats sont présentés ci-dessous.

4 Résultats et discussion

Les expérimentations ont été menées à partir de retours d'expérience d'opérations de restauration recensées dans un fichier Excel, où chaque ligne correspond à une opération de restauration, et inclue des informations de différents types. Les performances des modèles de langage, notamment Qwen2.5, Qwen2, Mistral, Llama3.1 et Llama3.2 ont été évaluées en fonction de leur capacité à extraire des connaissances pour peupler l'ontologie de domaine. Cette évaluation repose sur le calcul de la précision, du rappel et de la F-mesure, obtenus grâce à une comparaison manuelle des triplets extraits.

	zero-shot				one-shot			
	Précision	*Rappel*	*F-mesure*	*Temps moyen*	*Précision*	*Rappel*	*F-mesure*	*Temps moyen*
Mistral	0.75	0.90	0.82	46s	0.74	0.77	0.69	1mn 32s
Llama 3.1	0.84	0.90	0.87	44s	0.83	0.75	0.78	58s
Llama 3.2	0.73	0.85	0.78	**25s**	0.69	0.70	0.69	**43s**
Qwen 2	0.77	0.89	0.82	37s	0.69	0.65	0.66	1mn 40s
Qwen 2.5	**0.89**	**0.94**	**0.91**	1mn 27s	**0.93**	**0.89**	**0.90**	2mn 21s

TAB. 1 – *Résultat des différents LLMs sur 15 retours d'expériences de restauration*

Le tableau 1 présente les performances de différents modèles de langage pour l'extraction de triplets à partir de 15 textes de retours d'expériences de restauration, comprenant 10 textes en français issus du travail de synthèse effectué par Schmitt et Beisel (2015) et 5 textes en anglais sélectionnés à partir du site de l'US Forest Service, National Stream & Aquatic Ecology Center (2024).

Les résultats obtenus montrent clairement que les modèles de grande taille et les plus récents, tels que Qwen2.5, affichent des performances supérieures pour les deux types de prompts. En effet, le modèle Qwen2.5 :14b affiche des performances nettement supérieures avec le prompt zero-shot guidé par un template, atteignant une précision de 0.89, un rappel de 0.94 et une F-mesure de 0.91, bien que son temps d'exécution moyen soit relativement élevé (1 minute 27 secondes). Sa grande taille, avec 14 milliards de paramètres, lui permet de mieux

intégrer et exploiter les informations contextuelles du prompt, ce qui se traduit par une identification précise des triplets corrects tout en minimisant les erreurs et les omissions. En revanche, les modèles intermédiaires, comme Mistral :7b, Qwen2 :7b (7 milliards de paramètres chacun) et Llama3.1 :8b (8 milliards de paramètres), obtiennent des performances similaires avec une F-mesure de 0.82. Toutefois, Qwen :7b est légèrement plus rapide (37 secondes contre 44 secondes pour Llama3.1 :8b et 46 secondes pour Mistral :7b). La précision de Llama3.1 :8b (0.84) se démarque toutefois de manière significative par rapport à celle de Mistral :7b (0.75) et de Qwen2 :7b (0.77), justifiant en partie l'augmentation du nombre de paramètres malgré un temps d'exécution légèrement plus long. Enfin, Llama3.2 :3b, avec seulement 3 milliards de paramètres, présente des résultats plus modestes (F-mesure : 0.78) mais se distingue par un temps d'exécution remarquable (25 secondes). Cela le rend potentiellement adapté pour des tâches nécessitant des réponses rapides, bien que ses performances restent limitées pour des entrées complexes.

Avec le prompt one-shot, qui un prompt imbriqué incluant un exemple, les performances restent globalement similaires mais montrent une légère dégradation pour certains modèles. Notamment, Qwen2.5 :14b reste le plus performant (F-mesure : 0.90), bien que le temps d'exécution atteigne 2 minutes 21 secondes. Cela démontre encore la puissance de ce modèle pour intégrer des contextes complexes, notamment des prompts imbriqués avec un exemple fourni. Cependant les performances des modèles Qwen :7b, Mistral :7b et Llama3.1 :8b diminuent légèrement, avec des F-mesures respectives de 0.66, 0.69 et 0.78. La dégradation pour Qwen2 :7b et Mistral :7b pourrait être liée à une architecture moins adaptée à ce type de prompt. Toutefois, Llama3.1 :8b, avec ses 8 milliards de paramètres, montre une certaine robustesse en one-shot tout en maintenant un temps d'exécution raisonnable de 58 secondes. Les modèles plus petits, tels que Llama3.2 :3b (3 milliards de paramètres), montrent une baisse significative des performances avec le prompt one-shot, atteignant une F-mesure de 0.69. Cette diminution reflète la difficulté accrue pour ce type de modèle à intégrer et exploiter l'exemple fourni, en raison de sa taille réduite. Toutefois, il conserve un avantage notable avec un temps d'exécution moyen le plus rapide (43 secondes) par rapport aux autres modèles.

Le tableau 2 présente les résultats de génération des triplets avec les différents LLMs exploités à partir d'un exemple de retour d'expérience de restauration. Pour chaque modèle LLM, les réponses de chaque prompt sont montrées. Les triplets surlignés en jaune sont considérés comme des assertions erronées. Cette illustration visuelle confirme que le modèle Qwen2.5 :14b reste systématiquement le meilleur modèle, ce qui est cohérent avec les résultats quantitatifs présentés dans le tableau 1.

5 Conclusion

Dans cet article, nous avons proposé une modélisation ontologique des concepts de domaine de restauration des hydro-écosystèmes, à partir de données provenant des retours d'expérience de restauration. Par la suite, nous avons étudié la capacité des grands modèles de langages pour peupler cette ontologie de domaine. Les expériences, sur deux types de prompts, montrent des résultats satisfaisants, en particulier pour les modèles de grande taille (en nombre de paramètres). Ces résultats peuvent être améliorés en exploitant des versions plus grandes de ces modèles ou bien d'autres. Une perspective intéressante consisterait à faire le fine-tuning

d'un modèle LLM pour l'adapter spécifiquement aux particularités du domaine de la restauration des hydro-écosystèmes.

Texte d'entrée *:*
Mesure de restauration : Passe à poissons de Brisach. Masse(s) d'eau : Le Rhin
Catégorie : Continuité longitudinale. Pays : France. Département/Land : Haut-Rhin.
Commune(s) : Brisach. Maître d'ouvrage : CERGA SA Pression(s) sur la zone : Construction de la microcentrale hydroélectrique en 2008. Cette centrale franco-allemande (EDF et EnBW) utilise la hauteur de chute de 5,7 m crée par le barrage agricole construit de 1962 à 1965. Travaux et aménagements : Création de la passe à poissons, avec : • 4 entrées : 2 en surface et 2 en profondeur. • Une passe à poissons « classique » pour la montaison. • Système de dévalaison (2 ouvrages) installé de part et d'autre de la grille permettant aux poissons de contourner la turbine. • Dispositif de piégeage pour le suivi scientifique des poissons.
Suivi biologique : oui. Suivi hydromorphologique : non. Compartiment(s) suivi(s) : Ichtyologie

	Triplets extraits (Format Turtle)	
LLM	**Prompt Zero-Shot**	**Prompt One-Shot**
Qwen2.5	BrisachFishPass a :reconnection ; :associatedTo :BrisachDam ; :isManagedBy :CERGASA ; :hasMainContractor :"EDF and EnBW" ; :startsAt "2008-01-01T00:00:00Z"^^xsd:dateTime; :endsAt "2008-12-31T23:59:59Z"^^xsd:dateTime ; :hasMaxLowWaterFlow 100.0 ^^xsd:float ; :hasMinLowWaterFlow 20.0 ^^xsd:float ; :containsTechnicalElement :BrisachFishPassElements ; :numberOfPools 4 ; :averageSlope 5.7 ; skos:prefLabel "Passe à poissons de Brisach". :BrisachDam a :dam, TetraOnto:baffleBrush ; :isLocatedOn :BrisachArea ; :hasHeight 5.7 ; :isImpassable "true"^^xsd:boolean ; skos:prefLabel "Brisach Dam". :BrisachArea a :geographicZone ; skos:prefLabel "Brisach Area". :CERGASA a :projectOwner ; skos:prefLabel "CERGA SA". :"EDF and EnBW" a :mainContractor ; skos:prefLabel "EDF and EnBW". :BrisachFishPassElements a :Technical Element ; skos:prefLabel "Brisach Fish Pass Elements". :TetraOnto:hasBiologicalMonitoring "true"^^xsd:boolean ; :hasHydromorphologicaMonitoring "false"^^xsd:boolean.	:PasseAPoissonsDeBrisach a :successivesBasinsPass ; skos:prefLabel "Passe à poissons de Brisach" ; :isManagedBy :CERGASA ; :concerns :LongitudinalContinuity ; :hasBiologicalMonitoring "true"^^xsd:boolean ; :hasHydromorphologicalMonitoring "false"^^xsd:boolean ; :startsAt "2008-01-01"^^xsd:dateTime . :MicrocentraleHydroelectriqueDeBrisach a :dam ; skos:prefLabel "Microcentrale Hydroélectrique de Brisach" ; :isLocatedOn :LeRhin ; :hasHeight "5.7"^^xsd:decimal ; :isImpassable "true"^^xsd:boolean . :LeRhin a :river ; skos:prefLabel "Le Rhin" ; :isLocatedOn :HautRhinvBrisachFrance . :CERGASA a :mainContractor ; skos:prefLabel "CERGA SA" . :HautRhinvBrisachFrance a :geographicZone ; skos:prefLabel "Haut-Rhin, Brisach, France" .
Qwen2	:<RestorationMeasureInstance> a :reconnection; :associatedTo :river ; :isManagedBy :CERGA_SA; :hasMainContractor :Maître_d'ouvrage ; :startsAt "2008-01-01T00:00:00"^^xsd:dateTime; :endssAt "2013-12-31T23:59:59"^^xsd:dateTime ; :hasMaxLowWaterFlow 4.0^^xsd:float ; :hasMinLowWaterFlow 1.5^^xsd:float> ; :containsTechnicalElement :passe_a_poissons ; :numberOfPools null ; hasBiologicalMonitoring true ; hasHydromorphologicaMonitoring false ; skos:prefLabel "Passe à poissons de Brisach". :<StructureInstance> a :dam ; :isLocatedOn :Masse_d'eau ; :hasHeight 5.7 ; :isImpassable false ; skos:prefLabel "Barrage agricole". :<GeographicZoneInstance> a :geographicZone ; skos:prefLabel "Comune(s) de Brisach". :<ProjectOwnerInstance> a :projectOwner ; skos:prefLabel "CERGA SA". :<MainContractorInstance> a :mainContractor ; skos:prefLabel "Maître d'ouvrage". :<TechnicalElementInstancet> a :passe_a_poissons ; skos:prefLabel "Passe à poissons".	# The restoration measure: :PasseAPoissonsDeBrisach a :successivesBasinsPass ; skos:prefLabel "Passage pour poissons de Brisach" ; :associatedTo :LeRhin ; :isManagedBy :CERGA_SA ; :containsTechnicalElement :Puce_Suivi_Scientifique_Poissons ; :concerns Continuite_longitudinale ; :startsAt "2008"^^xsd:dateTime ; :numberOfBasins "4"^^xsd:int ; :endsAt "In progress" ; :hasCost xsd:string ; :hasBiologicalMonitoring "true"^^xsd:boolean ; :hasHydromorphologicalMonitoring "false"^^xsd:boolean ; :isManagedBy :CERGA_SA ; :concerns Continuite_longitudinale . # The river: :LeRhin a :river ; skos:prefLabel "Le Rhin" ; :isLocatedOn :Haut-RhinBrisach_France ; label : "Le Rhin" . # The project owner: :CERGA_SA a :projectOwner ; skos:prefLabel "CERGA SA" . # The technical element: :Puce_Suivi_Scientifique_Poissons a :technicalElement ; skos:prefLabel "Dispositif de piégeage pour le suivi scientifique des poissons" . # The geographic zone: :Haut-RhinBrisach_France a :geographicZone ; skos:prefLabel "Haut-Rhin, Brisach, France" .

(Suite)

Llama3.1	<RestorationMeasureInstance> a :reconnection ; :associatedTo :PasseAPoissonBrisachStructureInstance> ; :isManagedBy :CERGAProjectOwnerInstance> ; :hasMainContractor :CERGAMainContractorInstance> ; :startsAt "2008-01-01"^^xsd:dateTime; :endsAt "2010-12-31"^^xsd:dateTime ; :concerns :continuityLongitudinal ; :hasMaxLowWaterFlow "5.7"^^xsd:float> ; :hasMinLowWaterFlow "1.0"^^xsd:float> ; :containsTechnicalElement :PasseAPoissonBrisachTechnicalEl> ; :numberOfPools 4 ; :averageSlope 0.05; skos:prefLabel <"Passe à poissons de Brisach">. <PasseAPoissonBrisachStructureInstance> a :passage , :isLocatedOn :leRhineGeographicZoneInstance> ; :hasHeight "5.7"^^xsd:float> ; skos:prefLabel <"Passe à poissons de Brisach">. <leRhineGeographicZoneInstance> a :geographicZone ; skos:prefLabel <"Le Rhin">. <CERGAProjectOwnerInstance> a :projectOwner ; skos:prefLabel <"CERGA SA">. <CERGAMainContractorInstance> a :mainContractor ; skos:prefLabel <"CERGA SA">. < PasseAPoissonBrisachTechnicalEl > a :TechnicalElement , :hasPoolLength "10"^^xsd:float, :hasPoolWidth "5"^^xsd:float; skos:prefLabel <"Passe à poissons de Brisach technical element">.	:PasseAPoissonsDeBrisach a :successivesBasinsPass ; :associatedTo :MicrocentraleHydroelectriqueFrancoAllemande ; :isManagedBy :CERGASA ; :hasMainContractor :CERGA_SAMainContractor ; :numberOfEntrances "4"^^xsd:integer ; :hasCost 0.0^^xsd:decimal ; :hasBiologicalMonitoring "true"^^xsd:boolean ; :hasHydromorphologicalMonitoring "false"^^xsd:boolean ; :concerns :LongitudinalContinuity . :MicrocentraleHydroelectriqueFrancoAllemande a :dam ; :isLocatedOn :RhinBrisachFrance ; :isImpassable "true"^^xsd:boolean ; skos:prefLabel "Micro-centrale Hydroélectrique Franco-allemande" . :CERGASA a :mainContractor ; skos:prefLabel "CERGA SA" . :RhinBrisachFrance a :geographicZone ; skos:prefLabel "Le Rhin, Brisach, France" .
Llama3.2	:RestorationMeasureInstance a :RestorationMeasureClass ; :associatedTo : StructureInstance ; :isManagedBy : ProjectOwnerInstance ; :hasMainContractor : MainContractorInstance ; :startsAt "2008-01-01"^^xsd:dateTime ; :endsAt "2010-01-01"^^xsd:dateTime ; :hasMaxLowWaterFlow 5.7 ; :hasMinLowWaterFlow 4.2 ; :containsTechnicalElement :TechnicalElementInstance ; :numberOfPools 4 ; :numberOfBasins 1 ; :averageSlope 0.02 ; :skos:prefLabel "Passe à poissons de Brisach" . :StructureInstance a :Structure ; rdfs:subClassOf :basinPass ; :isLocatedOn :GeographicZoneInstance ; :hasHeight 5.7 ; :isImpassable true ; :skos:prefLabel "Barrage agricole" . :GeographicZoneInstance a :geographicZone ; rdfs:label "Rhin" . :ProjectOwnerInstance a :projectOwner ; rdfs:label "CERGA SA" . :MainContractorInstance a :mainContractor ; rdfs:label "CERGA SA" . :TechnicalElementInstance a :TechnicalElement ; rdfs:label "Système de dévalaison" ; :hasTechnicalElementDescription "Dévalaison pour les poissons" .	:PasseAPoissonsDeBrisach a :successivesBasinsPass skos:prefLabel "Passe à poissons de Brisach" ; :associatedTo skos:relatedTo BarrageAgricole ; :isManagedBy CERGASAComptentSuivicsu ; :concerns owl:equivalentTo :LongitudinalContinuity ; rdfs:label "Passe à poissons de Brisach" . :BarrageAgricole a :dam skos:prefLabel "Barrage Agricole" ; :isLocatedOn skos:relatedTo HautRhinBrisachFrance ; owl:disjointWith :geographicZone ; rdfs:label "Barrage Agricole" . :HautRhinBrisachFrance a :geographicZone skos:prefLabel "Haut-Rhin, Brisach, France" ; owl:equivalentTo DépartementduHautRhin ; rdfs:label "Haut-Rhin, Brisach, France" . :LeRhin a :river skos:prefLabel "Le Rhin" ; :isLocatedOn skos:relatedTo DépartementduHautRhin ; rdfs:label "Le Rhin" ; :CERGASAComptentSuivicsu a :otherActor skos:prefLabel "CERGA SA Compétent Suivi" ; owl:sameAs CERGASAComptentSuivicsu ; rdfs:label "CERGA SA Compétent Suivi" . :PasseAPoissonsDeBrisach has associatedTo :BarrageAgricole skos:prefLabel "Passe à poissons de Brisach" ; owl:sameAs CERGASAComptentSuivicsu ; rdfs:label "Passe à poissons de Brisach" .
Mistral	:Passe_a_poissons_de_Brisach a :restorationMeasure ; :associatedTo :Passe_a_poissons_de_Brisach_Structure ; :isManagedBy :CERGA_SA ; :hasMainContractor :CERGA_SA ; :startsAt "20XX-XX-XXT00:00:00"^^xsd:dateTime ; :endsAt "20XX-XX-XXT00:00:00"^^xsd:dateTime ; :containsTechnicalElement :Entrance1 ; :containsTechnicalElement :Entrance2 ; :containsTechnicalElement :Entrance3 ; :containsTechnicalElement :Entrance4 ; :containsTechnicalElement :DevalaisonSystem ; :containsTechnicalElement :Piégeage ; :numberOfPools 0 ; :averageSlope "N/A" ; skos:prefLabel "Passe à poissons de Brisach". :Passe_a_poissons_de_Brisach_Structure a :barrage, :passe_a_poissons ; :isLocatedOn :GeographicZone_Rhin ; :hasHeight "5.7" ; :isImpassable false ; skos:prefLabel "Passe à poissons de Brisach Structure". :GeographicZone_Rhin a :geographicZone ; skos:prefLabel "Le Rhin". :CERGA_SA a :projectOwner ; skos:prefLabel "CERGA SA".	:FishPassageBrisach a :artificialRiver ; skos:prefLabel "Fish Passage at Brisach, Rhine" ; :associatedTo :EtudeContinuiteLongitudinaleRhine ; :isManagedBy :EDF , :EnBW ; :hasMainContractor :MicroCentraleHydroelectriqueBrisach ; :startsAt "2008-12"^^xsd:dateTime; :ends "in progress"^^xsd:dateTime ; :hasCost "X € HT"^^xsd:string . :MicroCentraleHydroelectriqueBrisach a :hydroElectricPlant , :fishPassageStructure ; skos:prefLabel "Microcentrale Hydroélectrique de Brisach" ; :isLocatedOn :Rhine ; :associatedTo :FishPassageBrisach ; :hasHeight "5.7"^^xsd:decimal ; :averageSlope "X"^^xsd:decimal ; :isImpassable "false"^^xsd:boolean . [A1] a :fishPassageEntrance , :surfaceLevelEntrance ; skos:prefLabel "Entrée d'accès à la migration de poisson, niveau de surface" ; :isPartOf :MicroCentraleHydroelectriqueBrisach . [A2] a :fishPassageEntrance , :deepLevelEntrance ; skos:prefLabel "Entrée d'accès à la migration de poisson, niveau profond" ; :isPartOf :MicroCentraleHydroelectriqueBrisach . [B1] a :bypassStructure , :leftBypass ; skos:prefLabel "Structure de contournement (côté gauche)" ; :isPartOf :MicroCentraleHydroelectriqueBrisach .

TAB. 2 – Exemples d'extraction des triplets pour chaque LLM appliqué sur un exemple de retour d'expérience

Références

Aamodt, A. et E. Plaza (1994). Case-Based reasoning : Foundational Issues, Methodological Variations, and System Approachs. *AICOM 7*(1), 39–59.

Bai, J., S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, et al. (2023). Qwen technical report. arXiv :2309.16609.

Brown, T. B. (2020). Language models are few-shot learners. arXiv :2005.14165.

Kouhoue, J. M., M. Lefrançois, A. Lesage, J. Lonlac, A. Doniec, et S. Lecoeuche (2024). Ontologie de maintenance des bâtiments et capacités des Larges Modèles de Langage (LLM) pour le peuplement. In *35es Journées francophones d'Ingénierie des Connaissances*, pp. 103–106.

Li, B., G. Fang, Y. Yang, Q. Wang, W. Ye, W. Zhao, et S. Zhang (2023). Evaluating ChatGPT's Information Extraction Capabilities : An Assessment of Performance, Explainability, Calibration, and Faithfulness. arXiv :2304.11633.

Lubani, M., S. A. M. Noah, et R. Mahmud (2019). Ontology population : Approaches and design aspects. *Journal of Information Science 45*(4), 502–515.

OpenAI, R. et al. (2023). GPT-4 technical report. arXiv :2303.08774.

Polat, F., I. Tiddi, et P. Groth (2024). Testing prompt engineering methods for knowledge extraction from text. *Semantic Web. Under Review*.

Sahbi, A. N. E., C. Alec, et P. Beust (2024). Peuplement automatique d'ontologie : l'ia générative est-elle plus efficace qu'une approche sémantique ? In *24ème conférence francophone sur l'Extraction et la Gestion des Connaissances (EGC)*.

Schmitt, L. et J.-N. Beisel (2015). Les projets de restauration du Rhin Supérieur : vers la mise en place d'un observatoire transfrontalier et transdisciplinaire. In *Final Proc. Int. Conf. Integrative Sciences and Sustainable Development of Rivers*, Volume 3.

Touvron, H., L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. (2023). Llama 2 : Open foundation and fine-tuned chat models. arXiv :2307.09288.

US Forest Service, National Stream & Aquatic Ecology Center (2024). FishXing : Case Studies. https ://www.fs.usda.gov/biology/nsaec/fishxing/case.html.

Summary

Restoring hydrological ecosystems is crucial for maintaining the quality and sustainability of water resources. Managers of hydraulic infrastructures, such as dams or hydroelectric plants have a legal obligation to restore part of their domain in order to contribute to these efforts. Numerous restoration projects have been carried out, but the lack of systematic feedback limits the ability of managers to plan and execute new projects effectively. This article presents an innovative approach to structuring and utilizing these data through the automatic population of a domain ontology, supported by large language models (LLMs).

Multi-embedding d'un graphe de connaissances pour le repositionnement de médicaments - Retour d'expérience

Malika Smaïl-Tabbone
Université de Lorraine-LORIA/CNRS

Cette soumission est le résumé d'un article publié dans la revue *Scientific Reports* [Islam et al. (2023)]. J'y rapporte quelques leçons apprises et perspectives futures.

1 Introduction et Contexte

La pandémie de COVID-19 a suscité de nombreux efforts pour découvrir des traitements efficaces mais le développement de nouveaux médicaments reste long et coûteux. Le repositionnement de médicaments s'impose depuis quelques années comme une stratégie efficace en limitant l'exploration à des médicaments déjà approuvés, dont les profils de sécurité et pharmacologiques sont connus. Cependant, un défi persistant est de prédire efficacement l'efficacité des médicaments contre le SARS-CoV-2 sans multiplier les essais cliniques infructueux. Pour relever ce défi, nous utilisons un graphe de connaissances (GC) comme une base qui relie divers types d'informations biologiques et cliniques. Un tel GC intègre des données variées provenant de différentes sources, offrant une perspective plus large que le simple docking moléculaire. Nous exploitons un GC centré sur la COVID-19 pour prédire les interactions possibles entre les médicaments et les protéines virales du SARS-CoV-2 (virus responsable de la COVID-19). Ce graphe nous permet de capturer la complexité des relations biologiques dans une représentation vectorielle.

Notre but était de créer un cadre d'apprentissage hybride, associant un graphe de connaissances (GC) à un modèle d'apprentissage afin d'identifier des médicaments prometteurs tout en fournissant des explications compréhensibles sur les prédictions. En intégrant des embeddings dans un réseau neuronal, nous pouvons prédire en aval des molécules candidates tout en fournissant des justifications basées sur des chemins explicatifs dans le GC. Notre approche combine deux modèles d'embedding géométriques (TransE, TranH) et un modèle sémantique (DistMult). L'extraction de règles explicatives issues du GC enrichit également l'analyse en justifiant les prédictions. En plus de l'évaluation des performances du modèle de prédiction, nous validons les molécules prédites par docking moléculaire.

2 Méthodologie

Nous avons exploité un graphe de connaissances spécifique à la COVID-19 intégrant plusieurs bases de données (DrugBank, Hetionet, STRING, IntAct, etc.) et des publications sur le

SARS-CoV-2. DRKG [Ioannidis et al. (2020)]. La version nettoyée du graphe, contient environ 98 000 entités et près de 6 millions de relations. Chaque relation prend la forme de triplets (entité-objet, relation, entité-sujet), ce qui permet de structurer les interactions biologiques et moléculaires complexes avant leur analyse. La figure 1 montre les types d'entités et le nombre de relations distinctes.

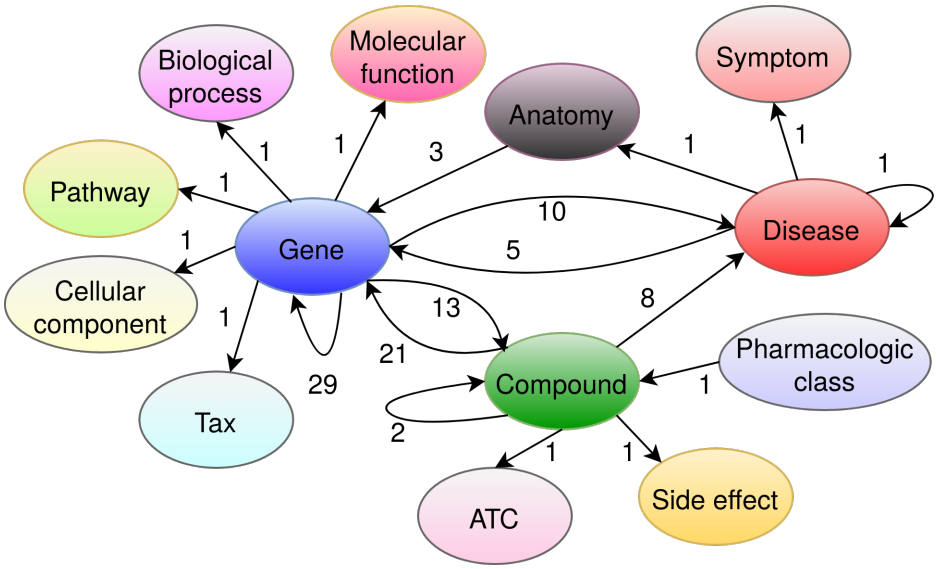


FIG. 1 – *Types d'entités présentes dans le GC DRKG ainsi que le nombre de relations distinctes reliant chaque paire de types d'entités.*

Pour apprendre à représenter chaque entité et relation dans le GC, nous appliquons trois méthodes d'embedding : TransE, TransH et DistMult, qui offrent une vue complémentaire des relations. Une analyse en composantes principales (PCA) est ensuite appliquée pour réduire la dimensionnalité. En combinant les embeddings des trois méthodes, nous obtenons des représentations finales d'entités et de relations de dimension 74, prêtes à être utilisées dans un réseau neuronal.

1. TransE (Translation Embedding) repose sur une approche de translation, où les relations entre les entités sont modélisées par des vecteurs de translation. Dans cette méthode, si une relation existe entre une entité source et une entité cible (par exemple, un médicament et une cible virale), le vecteur de l'entité source, additionné au vecteur de relation, devrait approximer le vecteur de l'entité cible.

2. TransH (Translation in Hyperplane) affine le modèle de TransE en introduisant un hyperplan spécifique pour chaque relation. Cette méthode décompose les entités selon l'hyperplan, facilitant ainsi la gestion de relations 1-N et N-1.

3. DistMult (Distance Multiplier) repose sur la similarité sémantique et modélise les interactions en calculant le produit matriciel des vecteurs des entités et des relations. Cette méthode utilise des poids scalaires pour évaluer les interactions.

Ces trois méthodes, une fois appliquées, permettent de capturer des relations complexes dans le graphe de connaissances.

Nous avons ensuite entraîné un réseau de neurones multi-couches utilisant des embeddings d'entités. En nous appuyant sur un ensemble de paires (médicament, cible), nous entraînons ce modèle à détecter les relations « Traite » entre une maladie (représentée par une protéine du virus) et une molécule. L'évaluation a montré des scores élevés de précision avec une erreur quadratique moyenne de 0,09 et une AUC de 0,96.

Nous avons mené une évaluation moléculaire ciblant la protéine nsp13, essentielle à la réplication du SARS-CoV-2. Deux méthodes complémentaires ont été utilisées : une évaluation par similarité des ligands et un docking moléculaire. Par clustering, nous avons identifié des groupes de structures similaires, révélant que plusieurs ligands prédits partagent des caractéristiques structurelles avec des ligands connus. Le docking moléculaire a permis de révéler le Fosinopril comme un ligand avec une forte affinité pour nsp13.

3 Analyse des résultats et perspectives

Sur le plan applicatif, notre modèle a identifié 100 candidats potentiels pour le traitement de la COVID-19, parmi lesquels des médicaments déjà en essais cliniques, confirmant la fiabilité de notre approche. L'analyse approfondie du Fosinopril a montré des scores de docking élevés, suggérant une forte affinité avec nsp13. Par ailleurs, la corrélation de similarité moléculaire renforce cette hypothèse, plaçant le Fosinopril comme un potentiel inhibiteur de la protéine virale. Les résultats pour la molécule Fosinopril sont prometteurs et justifient d'aller plus loin vers une validation expérimentale. Ainsi, notre étude montre comment l'utilisation combinée d'embeddings d'un GC et de docking moléculaire nous permet de proposer des médicaments candidats pour une pathologie particulière. Ce cadre pourrait être adapté à d'autres maladies pour peu que l'on dispose de graphes idoines.

Par rapport à l'explicabilité, en instanciant des règles dans le GC, nous avons pu tracer des chemins reliant les entités médicaments aux cibles virales. Par exemple, pour le Fosinopril, les chemins d'interaction incluent des intermédiaires pertinents, soulignant des connexions moléculaires et renforçant la confiance dans notre prédiction.

Sur le plan méthodologique, notre approche se démarque en intégrant plusieurs méthodes d'embeddings, une validation moléculaire et une analyse explicative ancrée sur le GC. Bien que notre approche ait donné des résultats prometteurs, nous avons constaté que la qualité des embeddings restait dépendante de celle des données contenues dans le GC, nous obligeant à nettoyer les données souvent bruitées et parfois contradictoires issues du moissonnage de sources de données de qualité inégale. De plus, les données sur la COVID-19 étant encore en constante évolution, une actualisation régulière du GC est nécessaire pour maintenir la pertinence des prédictions.

Les méthodes d'embeddings géométrique ou sémantique utilisées ont l'avantage de la simplicité mais leur mise en oeuvre reste lourde sur des grands graphes de connaissances (quelques centaines d'époques d'apprentissage nécessitent plusieurs jours de calcul). Nous avons également pu vérifier sur le graphe DRKG ce qui avait été pointé dans la littérature [Rossi et al. (2021)] à savoir que TransE est efficace pour modéliser les relations 1-1 mais a des limites avec les relations symétriques ou plus complexes, TransH s'avère plus efficace pour modéliser les

relations asymétriques et DistMult est efficace pour les relations symétriques mais l'est moins pour modéliser des relations anti-symétriques.

Ainsi notre principale perspective est de réduire la complexité de notre modèle tout en améliorant la qualité des représentations par l'adaptation de certaines méthodes GNN à base de transmission de messages [Schlichtkrull et al. (2018)] afin de prendre en compte la spécificité de certaines relations en fonction de leurs propriétés et de leur sémantique. L'intégration de mécanismes d'attention dans ces méthodes devrait nous permettre de gérer l'incertitude et le bruit inévitables dans les données mais demeurant difficiles à débusquer.

Références

Ioannidis, V. N., X. Song, S. Manchanda, M. Li, X. Pan, D. Zheng, X. Ning, X. Zeng, et G. Karypis (2020). Drkg-drug repurposing knowledge graph for covid-19. *GitHub* `https://github.com/gnn4dr/DRKG`. Last accessed 01 January 2022.

Islam, M. K., D. Amaya-Ramirez, B. Maigret, M.-D. Devignes, S. Aridhi, et M. Smaïl-Tabbone (2023). Molecular-evaluated and explainable drug repurposing for covid-19 using ensemble knowledge graph embedding. *Scientific Reports 13*(1), 3643.

Rossi, A., D. Barbosa, D. Firmani, A. Matinata, et P. Merialdo (2021). Knowledge graph embedding for link prediction : A comparative analysis. *ACM Transactions on Knowledge Discovery from Data (TKDD) 15*(2), 1–49, doi: 10.1145/3424672.

Schlichtkrull, M., T. N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, et M. Welling (2018). Modeling relational data with graph convolutional networks. In *The semantic web : 15th international conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, proceedings 15*, pp. 593–607. Springer.

A LLM centred framework for ontology generation in urban planning domain

Rania Bennetayeb*,**, Giuseppe Berio*, Nicolas Béchet*

{rania.bennetayeb,giuseppe.berio,nicolas.bechet}@univ-ubs.fr
* Univ. Bretagne Sud, IRISA, Vannes, France
**Rennes Métropole

Abstract. The growing complexity of urban data has driven the need for automated solutions to manage land use laws. This paper presents an LLM-based framework to generate ontologies for the urbanism domain by extracting structured triples from regulatory documents. The framework evaluates LLMs for triples extraction, extracts and validates triples to enrich a core ontology. Preliminary results with GPT-4 and Claude highlight challenges in capturing semantic relationships using exact matching. To address this, we plan to explore few-shot prompting and fine-tuning with semantic evaluation. The resulting ontology will enable efficient and reliable building permit analyses, adapting to evolving regulations.

1 Introduction

The increasing demand to automate tasks in various fields has driven the need for efficient methods of structuring data. Ontologies and Knowledge Graphs (KG) are powerful tools for formalizing domain-specific knowledge, enabling enhanced data integration, interoperability, and reasoning. These structures provide a semantic framework that facilitates human understanding and empowers machines to process complex information. Recent advancements in Natural Language Processing (NLP) have made ontology generation less labor-intensive, especially given the vast amounts of data that need to be processed to enable automation.

Large language models (LLMs) such as GPT (Achiam et al. (2023)), Mistral (Jiang et al. (2023)) and Llama (Touvron et al. (2023)), excel at extracting meaningful patterns and relationships from unstructured textual data, making them valuable for precise knowledge representation.

In this article, we propose an initial framework to generate an ontology for the urban planning domain by leveraging the capabilities of LLMs. Our approach focuses on extracting triples from textual data in regulatory documents, such as Local Urban Plan (LUP) [1]. Triples, in the form of (subject, relation, object) (Hogan et al., 2021), capture semantic relationships in texts. For example, "Paris is the capital of France" is

1. LUP: known in French as the “Plan Local d’Urbanisme (PLU)”. This document organizes urban development coherently by defining rules for development and land use within its territory, responding to local needs and adapting to changes over time.

represented as (Paris, CapitalOf, France). Each extracted triple is a fundamental node for constructing an ontology or KG.

However, building this ontology is challenging, especially for regulatory French texts, due to limited resources like pre-trained models or annotated datasets tailored to domain-specific contexts (Abu-Salih, 2021), like urban planning. These limitations necessitate a domain-adapted approach for knowledge extraction and formalization.

2 Motivation

Urban planning regulatory documents and related laws are critical for structuring territories and governing urban developments. In Rennes Métropole (RM) [2], the inter-municipal LUP (Morand-Deviller and Ferrari, 2023) provides the building and land-use rules for the 43 municipalities in the metropolitan area.

The LUP specifies urban and natural zones, building rules, land uses (e.g., housing, industrial activities, green spaces), and public facilities. It also sets out land occupation (Jacquot et al., 2022) and use conditions such as building heights, parking requirements, and environmental standards. These documents are complex, heterogeneous, interdependent, and continuously evolving, requiring expertise for proper interpretation.

Examining a buliding permit (BP) against these rules is challenging. Inspectors face difficulties due to the complexity of projects, some vagueness and ambiguities of the regulations, and a correct appreciation of the rigidity degree related to the rules. For instance, some rules are strict, while others serve as guideline. Moreover, multiple rules may apply to the same element (e.g., fences, parking), derived from overlapping regulatory documents or the project's location in multiple zones. These difficulties are compounded by tight deadlines and increased workloads during peak periods, requiring additional staff who may lack the necessary expertise.

Thus, our primary goal is to develop a tool to assist instructors and speeding up the overall process. We have suggested the main component of the system has to be a domain-specific ontology, enabling reasoning and queries to check BP compliance with regulations. Accordingly, three main challenges are identified :

1. Ontology Development i.e. how to represent regulatory documents in an ontology that identifies key concepts (including complex concepts requiring axioms) and relationships, independent of specific BPs.
2. BP Cross-referencing i.e how to align BP descriptive elements with the ontology to verify compliance through a detailed evaluation process.
3. Adaptability i.e. how to ensure the ontology and the system remain updated whenever regulatory documents evolve.

3 Related works

The process of ontology construction involves several key steps. Based on a comprehensive review of the literature (Tamašauskaitė and Groth, 2023). The construction

2. Rennes Métropole: A metropolitan area in the Ille-et-Vilaine department, Brittany, France.

follows a systematic process involving data identification, ontology creation, knowledge extraction, refinement, and ongoing maintenance. Data may originate from structured, semi-structured, or unstructured formats.

Information Extraction (IE),Information Extraction (IE), a core NLP task encompassing named entity recognition, relation extraction, and terminology extraction, forms the foundation for constructing ontologies and Knowledge Graphs (KG).

Traditional multi-stage IE pipeline, generally include sequential steps such entity recognition, entity disambiguation, and RE. While effective, they suffer from error propagation across stages. An innovative alternative is the unification of these steps into a single process, known as End-to-End IE (Josifoski et al., 2021).

Among the most advanced techniques in IE methodologies are Prompt Engineering (PE) and Fine-Tuning (FT). PE involves designing task-specific, context-aware prompts that guide LLMs in extracting structured data, such as triples, directly from unstructured text. This technique is especially valuable in scenarios with limited annotated data, where traditional supervised learning methods may not be feasible. It can be effectively utilized in Zero-Shot Learning (ZSL), One-Shot Learning (OSL), and Few-Shot Learning (FSL) settings, where the model is required to perform tasks with minimal or no task-specific examples. By providing contextually rich inputs to the model, PE enables it to extract relevant information with a higher degree of accuracy, despite limited training data. In contrast, FT involves adapting pre-trained LLMs to specific tasks using various types of datasets. One widely used method is Supervised FT, where the model's parameters are adjusted based on labeled training data, ensuring effective generalization to the triplet extraction task.

Together, these techniques represent a state-of-the-art approach to End-to-End Information Extraction, enabling LLMs to autonomously and efficiently extract complex relationships from large-scale, unstructured datasets, thereby making significant strides in the advancement of IE and KG construction.

4 Framework

The proposed framework provides a LLMs centred approach for ontology generation and knowledge structuring, validation processes, and application-driven assessment.

In the remainder the central notion of triple will be used. This is because the relevant literature shows that LLMs work with triples (S,R,O) establishing a semantic relationship R between a Subject S and an Object O. However, any triple can comprise concepts (mostly generic terms not referring to one individual or thing, such as terms 'building", "destination", "activity" in the urbanisation domain) or instances (terms referring to specific individuals or things, such as the names of specific suburbs, towns, specific building). Relationship R can be predefined or extracted : in the first case, the LLM is mainly used to assess the existence, in the second case the LLM suggests the relationship. It could be possible that both extraction and existence should be taken into account for managing potential redundancy and generalisation.For instance, given the sentence "The total area of building named le soleil is about 2330 m2", expected triples may be "le soleil" is a "building", and "le soleil, hastotalsurface, "2330m2"" and "2330m2" hasunit "m2" and "2330m2" hasvalue "2330". This expectation could be

generalised by providing the fixed relationships "is a", "hastotalsurface", "hasunit" and "hasvalue". Each step required by the approach within the framework is explained below.

Tool selection and validation for triple extraction. In the initial stage, we aim to identify the most efficient LLM for the triple extraction task. However, as no LUP-specific annotated data is available to evaluate the extracted triples, we rely on public WebNLG+2020 [3] annotated data, reference sentences aligned with the corresponding triples. Using LLMs identified in the literature, we perform triples extraction in a OSL mode and evaluate the results of each LLM against the reference dataset. The evaluation focuses on the accuracy, consistency and semantic alignment of the extracted triples with the ground truth of the dataset. Section 5 presents the results of the preliminary experiments.

Triple extraction from regulatory texts. Selected tool(s) is then used to extract triples from the LUP. The extraction process employs OSL, leveraging the model's contextual understanding to generate triples in a structured RDF compatible format. A usage of a core ontology (see step below) can be integrated in this step.

Validation of extracted triples. The extracted triples need to be validated before being used further. Classical validation can be performed involving annotators (mainly RM instructor clerks), who manually evaluate the correctness and relevance of the triples. This ensures an earlier data quality which is specifically required for the system to be developed. However, only a validation through the application will be able to assess the ontology quality.

Ontology Development and Enrichment. Extracted triples could be used for building from scratch a complete ontology. However, we think that this is not necessarily effective. A better way of working is to manually provide a quite thin ontology representing the main concepts, relationships and axioms (core ontology). The triples will be therefore used for enriching the core ontology by mainly incorporating additional concepts and relationships as well as to develop a KG related to the core or enriched (Kollapally, 2024). The given ontology can be used for contextualising the specific requests as for instance saying where given subject/object can be classified or asking which kind of relationship should be considered between subject and object. Obviously a continuous consistency checking has be performed to guarantee that the resulting ontology is consistent.

Validation through Application. The final validation step assesses the ontology through prototyping the real-world application. A selection of BPs is manually checked by instructors; the same is done by reasoning and/or querying the ontology once each BP is also represented by using the ontology. If both ways of checking BPs achieve the same results than the ontology can be considered as valid for the application on hand.Otherwise a revision of the ontology should be undertaken.

3. `https://huggingface.co/datasets/GEM/web_nlg`

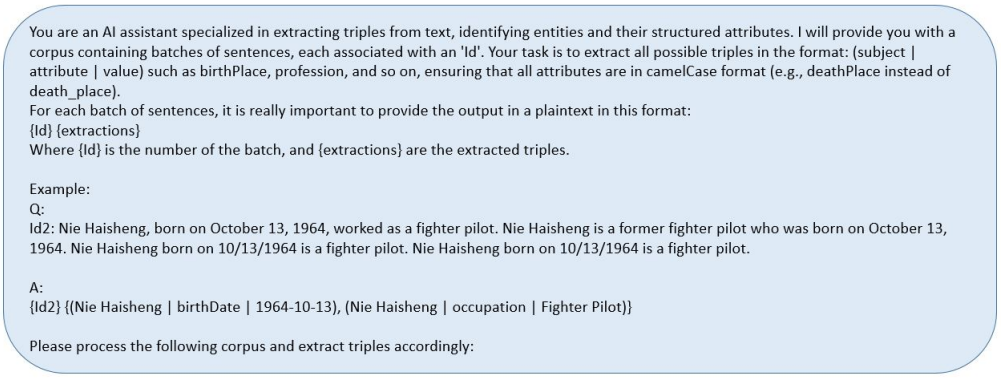


Fig. 1 – *Prompt*

5 Preliminary work

Subsection 5.1 focuses on triple extraction with LLMs and and presents performance figures related to two state-of-the-art LLMs. Established metrics are employed to evaluate their performance. Subsection 5.2 discusses earlier attempts to develop the core ontology.

5.1 LLMs for triples extraction

To extract the triples, we used a subset of 150 first sentence of the WebNLG+2020 test dataset. We performed OSL with the GPT-4o and Claude 3.5 Sonnet models using an adapted prompt illustrated in figure 1, inspired by Ghanem and Cruz (2024).

Table 1 presents a relatively low performance in aligning extracted triples with reference triples, both in terms of exactness and coverage. Similarly, Claude 3.5 achieves slightly lower scores indicating greater difficulty in achieving exact matches. These results emphasize the challenges faced by both models in capturing accurate relationships and exact entities.

The divergence between strict extraction metrics and textual similarity metrics, reveals complementary strengths and weaknesses in both models. Precision, Recall, and F1-Score focus on exact matches, penalizing models heavily for minor deviations in generated triples. While, BLEU and ROUGE scores reward broader textual relevance and coherence, offering a more lenient perspective. This contrast suggests that both Claude 3.5 and GPT-4o struggle with fully capturing complex semantic triples, but they excel in producing contextually appropriate and linguistically relevant outputs.

5.2 Ontology development attempts

We have manually developed a earlier core ontology based on a surface analysis of key documents (specifically the LUP) and analysis of other ontologies for the same domain (Fauth and Seiß (2023)). This core ontology is needed for building the domain

	Claude 3.5	GPT-4o
Precision	0.31	0.33
Recall	0.30	0.32
F1 score	0.30	0.32
BLEU	0.55	0.58
ROUGE-1	0.80	0.81
ROUGE-L	0.66	0.66

TAB. 1 – *Perfomance evaluation of triples extraction*

ontology by systematic enrichment. The core ontology is also useful for understanding from the beginning the potential capability to answer specific competency questions on BPs. For instance, we can ask if a case BP is a instance of a specific concept, being the BP represented according to its characteristics such as land, surface, number of floors, wall materials, employed colors. Trivially, adding a BP represented as such to the ontology, the question may be answered by checking the ontology consistency. Concepts of this earlier core ontology are shown in figure 2.The core ontology is currently provided in Protégé [4]. By the way concepts are labelled in French but the essential structure is language independent.

The manual development of the core ontology is complemented by a deeper analysis of the key documents. Basically, the documents have been cleaned and most frequent terms have been extracted. The analysis can be useful to evaluate any manual enrichment of the earlier core ontology. For instance, some frequent terms may originate additional concepts because being frequent most of the BPs may refer to them.

Attempts presented above are required (and additional ones are still required) for starting the central task to fully generate a domain ontology showing that LLMs can take the central role of providing controlled enrichment of the earlier core ontology as also highlighted in relevant literature (e.g. Mihindukulasooriya et al. (2023)).

6 Future works

In our preliminary experiments, a OSL approach was utilized to guide the models in generating triples. However, future work will focus on a few-shot prompting strategy, where the process will involve analyzing the model's outputs to identify patterns or structures it struggles to generate accurately. Based on this analysis, we will design prompts featuring more diverse and challenging examples, such as unit measurements, date formats, and other complex structures. These tailored examples aim to guide the LLM by providing clearer patterns and context, ultimately improving its ability to generate triples that are both semantically accurate and structurally faithful to the ground truth.

Furthermore, we intend to enhance our extraction process by FT an LLM for this task. According to the state of the art, this approach would enable the model to

4. https://protege.stanford.edu/

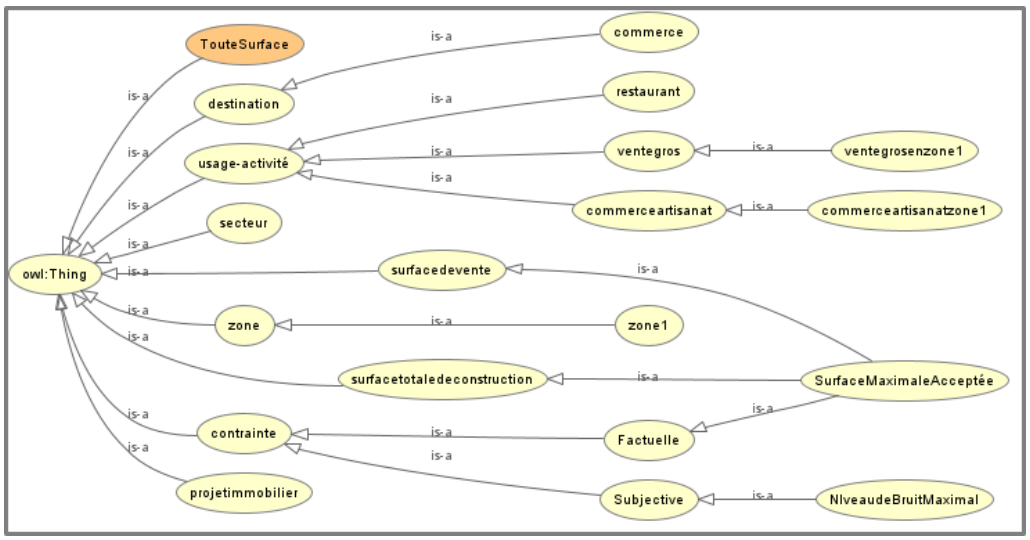


FIG. 2 – *Taxonomy of core ontology concepts*

better grasp the context and nuances of the task, thereby improving the accuracy and relevance of the extracted triples. For evaluation, we plan to integrate advanced methodologies such as G-Eval Liu et al. (2023), a methodology leveraging GPT-4 with chain-of-thought reasoning, to address the limitations of traditional metrics like BLEU and ROUGE, which often lack alignment with human judgment.

Concerning the work for developing a domain ontology, the current core ontology is continuously updated and will be used for contextualising triple extraction and for getting suggested enrichment from extracted triples. According to the framework, some work will be devoted to identify how to keep ontology consistency whenever suggested enrichment is done.

References

Abu-Salih, B. (2021). Domain-specific knowledge graphs: A survey. *Journal of Network and Computer Applications 185*, 103076.

Achiam, J., S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, and S. e. a. Anadkat (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Fauth, J. and S. Seiß (2023). Ontology for building permit authorities (obpa) for advanced building permit processes. *Advanced Engineering Informatics 58*, 102216.

Ghanem, H. and C. Cruz (2024). Fine-tuning vs. prompting: Evaluating the knowledge graph construction with llms. In *3rd International Workshop on Knowledge Graph*

Generation from Text (Text2KG), Co-located with the Extended Semantic Web Conference (ESWC 2024), May, pp. 26–30.

Hogan, A., E. Blomqvist, M. Cochez, C. d'Amato, G. De Melo, C. Gutierrez, and S. e. a. Kirrane (2021). Knowledge graphs. *ACM Computing Surveys (CSUR) 54*(4), 1–37.

Jacquot, H., F. Priet, and S. Marie (2022). *Droit de l'urbanisme* (9 ed.). Précis. Dalloz.

Jiang, A. Q., A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, and L. e. a. Saulnier (2023). Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Josifoski, M., N. De Cao, M. Peyrard, F. Petroni, and R. West (2021). GenIE: Generative information extraction. *arXiv preprint arXiv:2112.08340*.

Kollapally, N. M. (2024). *A Methodological Framework for Ontology Development, Enrichment, and Application in Natural Language Processing Tasks*. Ph. D. thesis, New Jersey Institute of Technology.

Liu, Y., D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu (2023). G-eval: NLG evaluation using gpt-4 with better human alignment. In H. Bouamor, J. Pino, and K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, pp. 2511–2522. ACL.

Mihindukulasooriya, N., S. Tiwari, C. F. Enguix, and K. Lata (2023). Text2kgbench: A benchmark for ontology-driven knowledge graph generation from text. In *International Semantic Web Conference*, pp. 247–265. Springer.

Morand-Deviller, J. and S. Ferrari (2023). *Droit de l'urbanisme* (11 ed.). Mémentos. Dalloz.

Tamašauskaitė, G. and P. Groth (2023). Defining a knowledge graph development process through a systematic review. *ACM Transactions on Software Engineering and Methodology 32*(1), 1–40.

Touvron, H., T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, and B. e. a. Rozière (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Résumé

La complexité croissante des données urbaines a engendré un besoin de solutions automatisées pour gérer les lois des droits des sols. Cet article présente un cadre basé sur les grands modèles de langage (LLM) pour générer des ontologies dans le domaine de l'urbanisme en extrayant des triples structurés à partir de documents réglementaires. Le cadre évalue les LLM pour l'extraction de triples, extrait et valide des triples afin d'enrichir une ontologie de base. Les résultats préliminaires obtenus avec GPT-4 et Claude mettent en évidence les défis liés à la capture des relations sémantiques en utilisant une correspondance exacte. Pour y remédier, nous prévoyons d'explorer l'ingénierie de prompt à quelques exemples et le Fine-tuning avec une évaluation sémantique. L'ontologie ainsi obtenue permettra des analyses efficaces et fiables des permis de construire, en s'adaptant à l'évolution des réglementations.

G-T2KG : approche indépendante du domaine, basée sur OpenIE et LLMs pour l'apprentissage de graphes de connaissances

Othmane Kabal*, Mounira Harzallah*
Fabrice Guillet*, Ryutaro Ichise**

*Nantes Université, CNRS, LS2N, UMR 6004,
F-44000 Nantes, France
nom.prenom@LS2N.fr
**Tokyo Institute of Technology, Tokyo, Japan
ichise@iee.e.titech.ac.jp

Résumé. Cet article présente G-T2KG, une approche innovante pour la construction automatique de graphes de connaissances (KG) à partir de textes. Cette approche pipeline, indépendante du domaine, contribue à réduire les erreurs dans le graphe de connaissances généré. Elle combine un système OpenIE, un nettoyage basé sur les syntagmes nominaux et une validation par LLMs. Elle est évaluée sur deux domaines (informatique et musique). Les résultats montrent une amélioration du rappel et une meilleure précision.

1 Introduction

Les graphes de connaissances (KG) sont essentiels dans le web sémantique, avec des applications variées comme les systèmes de questions-réponses et les moteurs de recommandation. Cet article aborde la construction automatique de KG à partir de textes bruts, comprenant l'extraction des triplets ainsi que le mapping des entités et des prédicats. Bien que facilitée par les récentes avancées en traitement du langage naturel, la construction automatique de KG reste limitée par la complexité du langage et la spécificité des domaines. Ainsi, certaines approches se concentrent sur la construction de KGs spécifiques à un domaine, en spécialisant les étapes de la reconnaissance d'entités nommées (NER) et d'extraction de relations (RE), mais elles souffrent souvent d'une réutilisabilité limitée et nécessitent des données annotées Abu-Salih (2021). A contrario, d'autres approches utilisant un système d'Open Information Extraction (OpenIE) extraient des triplets indépendants du domaine, mais génèrent aussi des triplets non pertinents, nécessitant un post-filtrage, ce qui compromet leur indépendance au domaine. De plus, les grands modèles de langage (LLMs) comme GPT-4 Achiam et al. (2023) et Llama Touvron et al. (2023), bien qu'indépendants du domaine, se trompent et fabulent, ce réduit la qualité des KGs générés. Pour surmonter ces limitations, nous proposons une approche hybride intitulé "General Text to KG" (G-T2KG), qui combine 3 aspects : un système OpenIE, un nettoyage basé sur les groupes nominaux et une validation par les LLMs. Dans la section 2, nous présentons notre approche et ses étapes. La section 3 est consacrée à son évaluation et

sa comparaison à d'autres approches en utilisant deux benchmarks, avant de conclure. Il est à noter que cet article est un résumé de Kabal et al. (2024) publié dans KES2024 ainsi que le code source et les benchmarks créés dans le cadre de cette recherche sont mis à la disposition de la communauté via GitHub [1].

2 Approche G-T2KG

G-T2KG est un pipeline structuré en cinq étapes (Figure 1). L'étape de prétraitement du texte (E1) prépare les corpus en nettoyant les données, en résolvant les coréférences et en segmentant les phrases afin de faciliter leur traitement par les étapes suivantes. L'étape d'extraction d'information (E2) s'appuie sur deux outils complémentaires. Le premier (E2.1) un modèle avancé du système Open Information Extraction (OpenIE 6) Kolluru et al. (2020), offrant un équilibre optimal entre rapidité et précision dans l'extraction des triplets, indépendamment du domaine, tout en disposant d'un analyseur avancé des coordinations, capable d'extraire plusieurs triplets à partir de phrases conjonctives en les décomposant, afin d'éviter toute perte d'information, ce qui améliore le rappel.

Le deuxième est l'Extracteur de Relations d'Hyperonymes (E2.2), basé sur des patrons lexicaux-syntaxiques, qui enrichit le graphe en identifiant des relations d'hyperonymie ("est-un(e)") non extraites par le premier outil Roller et al. (2018). L'étape de post-traitement (E3) se concentre sur le nettoyage et l'intégration des triplets extraits par l'OpenIE, qui sont souvent trop longs et contiennent des parties inutiles telles que les adjectifs et les prépositions, visant à identifier la plus petite phrase nominale significative pour les entités et en rectifiant les prédicats, tout en s'appuyant sur des règles syntaxiques (E3.1), afin de garantir une structure cohérente et exploitable des données. Elle est suivie par l'étape de validation binaire des triplets (E4), qui consiste à filtrer les triplets incorrects résultant du processus d'extraction d'information tout en préservant l'indépendance de l'approche vis-à-vis du domaine grâce à l'utilisation du LLM GPT-4 avec une technique de prompting zero-shot, évitant ainsi le besoin d'exemples, et garantissant que chaque triplet reflète fidèlement le sens de la phrase source.

Enfin, l'étape de mapping (E5), qui inclut le mapping des entités (E5.1) et des prédicats (E5.2). L'approche de mapping proposée par SciCeroDessí et al. (2022) est adoptée dans le pipeline.

3 Evaluation

Benchmarks. Étant donné que notre approche est indépendante du domaine, nous l'évaluons avec deux benchmarks (B) issus de domaines distincts. Le premier, en informatique (CS-B), nous permet de comparer notre approche à d'autres approches adaptes à ce domaine, telles que SciCero Dessí et al. (2022), ainsi qu'aux LLM (GPT-4). Le second, en musique (Music-B), permet d'évaluer l'adaptabilité de notre approche à des domaines variés. Pour créer le CS-B, nous avons sélectionné 12 résumés d'articles en informatique provenant du jeu de données Web of Science Kowsari et al. (2017), comprenant 108 phrases. Ces résumés ont été segmentés en phrases, et des triplets ont été extraits. Trois annotateurs ont examiné les triplets et décidé de leur inclusion dans le GS-B sur la base d'un vote majoritaire, ce qui a abouti à un

1. https://github.com/OthmaneKabal/G-T2KG

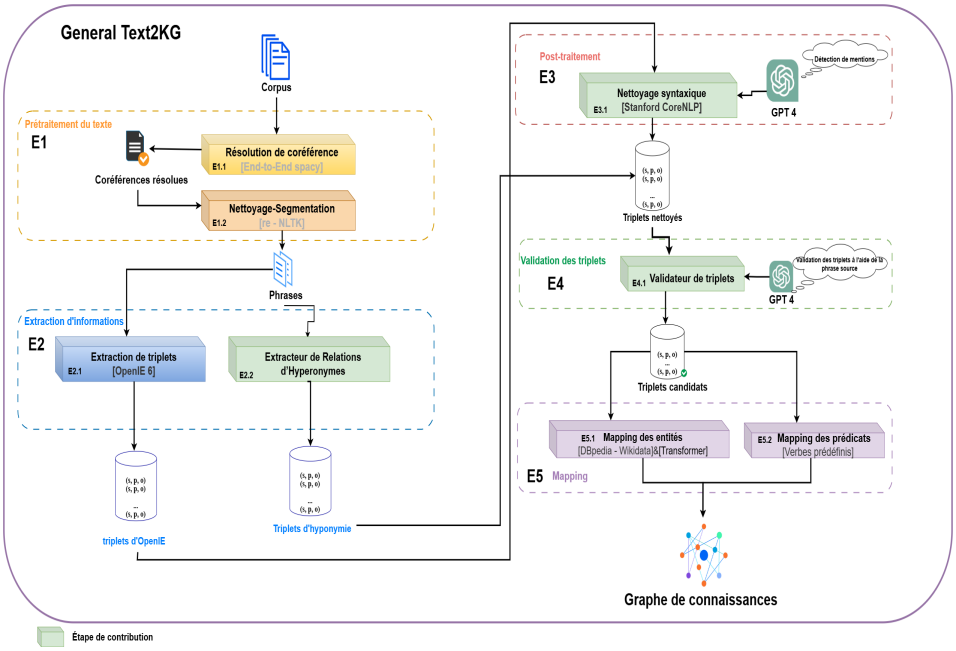


FIG. 1 – *Architecture General-T2KG.*

CS-B comprenant 180 triplets, 247 entités uniques et 100 prédicats uniques. Le Music-B a été dérivé du corpus fourni par Camacho-Collados et al. (2018), en sélectionnant aléatoirement le document "20th Century Music" avec 100 phrases. En appliquant la même méthodologie, nous avons abouti à un Music-B comprenant 410 triplets, avec 488 entités uniques et 155 prédicats uniques.

Evaluation. Pour évaluer notre approche, nous avons procédé à un appariement manuel des triplets extraits avec ceux du benchmark correspondant, en utilisant un vote majoritaire pour chaque triplet. Des métriques classiques (Précision, Rappel, F1) ont été utilisées pour la comparaison. Pour analyser l'impact de l'étape E4 (Validation des triplets), nous avons reconduit l'évaluation après sa suppression du pipeline G-T2KG. Nous avons également suivi la même procédure pour SciCero et mené deux expériences avec ChatGPT-4. La première (GPT-4-Exp 1) utilisait un prompting avec quelques exemples, demandant au modèle d'extraire des triplets d'une phrase. La seconde (GPT-4-Exp 2) apportait des instructions détaillées basées sur l'analyse des erreurs des résultats de la première, notamment en ajoutant des instructions et des exemples sur la gestion des coréférences ainsi que sur l'extraction de multiples triplets dans les cas de conjonctions. Les résultats sont illustrés dans les tables 1 et 2.

Les résultats de la table 1 montrent que G-T2KG surpasse les méthodes spécifiques au domaine, comme SciCero, en rappel pour le corpus CS. SciCero, limité par son approche ontologique, extrait des triplets spécifiques au domaine, ce qui augmente sa précision mais réduit son rappel. G-T2KG capture 23 % des triplets extraits par SciCero et détecte également des triplets manqués, comme *(network administrator ; prioritize ; vulnerability)*.

Corpus	Approches	Précision	Rappel	F1-score
Computer Science	G_T2KG_Opt1 sans TV	58.50%	**47.77%**	52.59%
	G_T2KG_Opt1	72.07%	44.44%	**54.98%**
	G_T2KG_Opt2 sans TV	50.00%	44.44%	47.05%
	G_T2KG_Opt2	**75.00%**	41.66%	53.56%
	SciCero	72.09%	17.22%	27.50%
	GPT-4-Exp 1	25.65%	32.77%	28.77%
	GPT-4-Exp 2	27.81%	43.88%	34.04%

TAB. 1 – *Résultats de l'évaluation de G-T2KG sur le corpus en informatique.*

Corpus	Approches	Précision	Rappel	F1-score
Music	G_T2KG_Opt1 sans TV	47.75%	36.34%	41.27%
	G_T2KG_Opt1	**64.88%**	35.12%	45.57%
	G_T2KG_Opt2 sans TV	46.97%	**37.56%**	41.74%
	G_T2KG_Opt2	63.97%	36.60%	**46.56%**

TAB. 2 – *Résultats d'évaluation de G-T2KG sur le corpus sur la musique.*

Comparé aux triplets générés par ChatGPT dans les expériences 1 et 2, G-T2KG excelle également en rappel et en précision. Les erreurs fréquentes de ChatGPT incluent une mauvaise gestion des conjonctions et des segmentations incohérentes *(tutorials for computer programming ; can be ; tedious to create)*. Ces observations confirment les conclusions de Zhu et al. (2023) sur les faiblesses de GPT-4 dans les corpus spécifiques.

La table 2 démontre l'applicabilité de G-T2KG à d'autres domaines, ici la musique, avec des performances constantes.

Enfin, l'intégration de GPT-4 comme outil de validation augmente significativement la précision (+14 à +25 %) avec un impact minimal sur le rappel (-3,3 % max), garantissant une construction de graphes de connaissances de qualité et adaptée à différents domaines.

4 Conclusion

Dans cet article, nous avons présenté l'approche G-T2KG pour la construction de graphes de connaissances à partir de textes, d'une façon indépendante du domaine. En combinant un système OpenIE et un système d'extraction des relations"est-un(e)", ainsi qu'en intégrant un nettoyage syntaxique et la validation des triplets via GPT-4, notre approche présente des performances supérieures en termes de précision. Les résultats obtenus soulignent l'intérêt des LLMs pour valider les triplets extraits.

Dans nos travaux futurs, nous prévoyons d'étendre l'utilisation des LLMs à d'autres étapes du processus de construction des graphes.

Références

Abu-Salih, B. (2021). Domain-specific knowledge graphs : A survey. *Journal of Network and Computer Applications 185*, 103076.

Achiam, J., S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv :2303.08774*.

Camacho-Collados, J., C. Delli Bovi, L. Espinosa-Anke, S. Oramas, T. Pasini, E. Santus, V. Shwartz, R. Navigli, et H. Saggion (2018). SemEval-2018 task 9 : Hypernym discovery.

Dessí, D., F. Osborne, D. R. Recupero, D. Buscaldi, et E. Motta (2022). Scicero : A deep learning and nlp approach for generating scientific knowledge graphs in the computer science domain. *Knowledge-Based Systems 258*, 109945.

Kabal, O., M. Harzallah, F. Guillet, et R. Ichise (2024). Enhancing domain-independent knowledge graph construction through openie cleaning and llms validation. *Procedia Computer Science 246*, 2617–2626, doi: https://doi.org/10.1016/j.procs.2024.09.436. 28th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES 2024).

Kolluru, K., V. Adlakha, S. Aggarwal, Mausam, et S. Chakrabarti (2020). OpenIE6 : Iterative Grid Labeling and Coordination Analysis for Open Information Extraction. In B. Webber, T. Cohn, Y. He, et Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, pp. 3748–3761. Association for Computational Linguistics, doi: 10.18653/v1/2020.emnlp-main.306.

Kowsari, K., D. E. Brown, M. Heidarysafa, K. J. Meimandi, M. S. Gerber, et L. E. Barnes (2017). Hdltex : Hierarchical deep learning for text classification. *CoRR abs/1709.08267*.

Roller, S., D. Kiela, et M. Nickel (2018). Hearst patterns revisited : Automatic hypernym detection from large text corpora. In I. Gurevych et Y. Miyao (Eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, Melbourne, Australia, pp. 358–363. Association for Computational Linguistics, doi: 10.18653/v1/P18-2057.

Touvron, H., L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. Canton Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. Singh Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, et T. Scialom (2023). Llama 2 : Open Foundation and Fine-Tuned Chat Models. *arXiv e-prints*, arXiv :2307.09288, doi: 10.48550/arXiv.2307.09288.

Zhu, Y., X. Wang, J. Chen, S. Qiao, Y. Ou, Y. Yao, S. Deng, H. Chen, et N. Zhang (2023). Llms for knowledge graph construction and reasoning : Recent capabilities and future opportunities. *ArXiv abs/2305.13168*.

Références

Abu-Salih, B. (2021). Domain-specific knowledge graphs : A survey. *Journal of Network and Computer Applications 185*, 103076.

Achiam, J., S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv :2303.08774*.

Camacho-Collados, J., C. Delli Bovi, L. Espinosa-Anke, S. Oramas, T. Pasini, E. Santus, V. Shwartz, R. Navigli, et H. Saggion (2018). SemEval-2018 task 9 : Hypernym discovery.

Dessí, D., F. Osborne, D. R. Recupero, D. Buscaldi, et E. Motta (2022). Scicero : A deep learning and nlp approach for generating scientific knowledge graphs in the computer science domain. *Knowledge-Based Systems 258*, 109945.

Kabal, O., M. Harzallah, F. Guillet, et R. Ichise (2024). Enhancing domain-independent knowledge graph construction through openie cleaning and llms validation. *Procedia Computer Science 246*, 2617–2626, doi: https://doi.org/10.1016/j.procs.2024.09.436. 28th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES 2024).

Kolluru, K., V. Adlakha, S. Aggarwal, Mausam, et S. Chakrabarti (2020). OpenIE6 : Iterative Grid Labeling and Coordination Analysis for Open Information Extraction. In B. Webber, T. Cohn, Y. He, et Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, pp. 3748–3761. Association for Computational Linguistics, doi: 10.18653/v1/2020.emnlp-main.306.

Kowsari, K., D. E. Brown, M. Heidarysafa, K. J. Meimandi, M. S. Gerber, et L. E. Barnes (2017). Hdltex : Hierarchical deep learning for text classification. *CoRR abs/1709.08267*.

Roller, S., D. Kiela, et M. Nickel (2018). Hearst patterns revisited : Automatic hypernym detection from large text corpora. In I. Gurevych et Y. Miyao (Eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, Melbourne, Australia, pp. 358–363. Association for Computational Linguistics, doi: 10.18653/v1/P18-2057.

Touvron, H., L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. Canton Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. Singh Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, et T. Scialom (2023). Llama 2 : Open Foundation and Fine-Tuned Chat Models. *arXiv e-prints*, arXiv :2307.09288, doi: 10.48550/arXiv.2307.09288.

Zhu, Y., X. Wang, J. Chen, S. Qiao, Y. Ou, Y. Yao, S. Deng, H. Chen, et N. Zhang (2023). Llms for knowledge graph construction and reasoning : Recent capabilities and future opportunities. *ArXiv abs/2305.13168*.

Leveraging LLMs for Content-Based Reviewer Assignment in Computer Science Conferences

Farid Bagheri*, Davide Buscaldi**, Diego Reforgiato Recupero*,***

*Department of Mathematics and Computer Science,
University of Cagliari, Via Ospedale 72, 09042, Cagliari, Italy
farid.bagheri@unica.it, diego.reforgiato@unica.it
http://www.unica.it
**Laboratoire d'Informatique de Paris Nord,Sorbonne Paris Nord University,
99 Av. Jean Baptiste Clement, 93430, Villetaneuse, France
davide.buscaldi@lipn.univ-paris13.fr
http://www.univ-paris13.fr
***Department of Mathematics and Computer Science,
University of Cagliari, Via Ospedale 72, 09042, Cagliari, Italy
diego.reforgiato@unica.it

Résumé. In this paper we investigate how Large Language Models (LLMs) and Knowledge graphs can be used effectively for content-based reviewer assignment in conference management systems. The assignment of appropriate reviewers to research papers is paramount in the organization of conferences and workshops. For this task, we employed a dataset composed of 663 papers from 85 conferences and 524 reviewer profiles, in the Semantic Web and Computer Science domain. By using a Large Language Model we identified keywords for each author or reviewer. Various similarity measures and representation strategies of these keywords are tested to evaluate the reviewer's relevance to each author's paper. The experiments conducted validate the effectiveness of our methodology for assigning reviewers to papers.

1 Introduction and Literature Review

The reviewer assignment problem is a crucial aspect of the academic peer review process, influencing the quality and fairness of scholarly evaluations. Effective reviewer matching ensures papers are assessed by experts with relevant expertise, fostering constructive feedback. Traditional methods, such as manual selection and basic keyword matching, were time-consuming, biased, and lacked scalability, often resulting in suboptimal reviewer-paper matches. Advances in NLP, machine learning, and optimization algorithms have since driven the development of automated and more efficient approaches to address these challenges.

To overcome the limitations of traditional reviewer assignment methods, researchers have developed advanced techniques leveraging NLP, machine learning, and op-

timization algorithms. For instance, Latypova (2023) introduced a methodology that evaluates reviewers not only by expertise but also by authority and the quality of past reviews, using text mining and the Analytic Hierarchy Process (AHP) to ensure a comprehensive and balanced match. Similarly, Abduljaleel et al. (2021) combined the Hungarian Algorithm with TF-IDF and cosine similarity for efficient matching, optimizing assignments for accuracy and fairness.

Hoang et al. (2021) proposed a decision support framework utilizing TF-IDF and Latent Dirichlet Allocation (LDA) for topic modeling, incorporating institutional diversity to reduce bias and improve assignment precision. Tan et al. (2021) introduced the Word and Semantic-based Iterative Model (WSIM), a hybrid approach integrating word-based and semantic features, demonstrating superior performance in matching accuracy through iterative modeling and advanced ranking techniques.

Arabzadeh et al. (2024) presented "Reviewerly", an IR-based framework treating submitted papers as queries and reviewers' past work as documents, utilizing databases like OpenAlex and fine-tuned language models such as SciBERT. Reviewerly also integrates diversity metrics to mitigate biases while ensuring a scalable, systematic reviewer assignment process.

Optimization techniques further enhanced these methods. Payan et Zick (2021) introduced algorithms like Greedy Reviewer Round Robin (GRRR) and FairSequence to address fairness and efficiency in large conferences, providing equitable distribution of reviews. Xu et al. (2022) combined coverage-based assignment with pre-trained language models and the Toronto Paper Matching System (TPMS), ensuring interdisciplinary coverage and incorporating conflict of interest detection for impartial reviews.

Building on these advancements, the integration of large language models (LLMs) has opened new pathways for enhancing the peer review process. Tyser et al. (2024) introduced innovative LLM-based systems, such as a framework for evaluating AI-generated reviews and the OpenReviewer tool, which leverages GPT-4 to produce consistent and transparent feedback. These systems address challenges like overconfident scoring, citation inaccuracies, and the need for adaptive review methods, underscoring both the potential and limitations of LLMs in automating reviews.

Further contributions include AGENTREVIEW, a simulation framework by Jin et al. (2024) that uses LLM agents to model reviewer dynamics and biases, and a gold standard dataset by Stelmakh et al. (2023), designed to benchmark similarity algorithms. However, these advancements also expose vulnerabilities, such as adversarial attacks on keyword-based systems, emphasizing the need for robust solutions. Studies have also explored fairness(Payan et Zick (2021)), diversity(Hoang et al. (2021); Arabzadeh et al. (2024)), and conflict of interest detection(Zhao et Zhang (2022)), as seen in methods incorporating institutional diversity and COI mechanisms(Xu et al. (2022)).

This evolving landscape underscores the ongoing efforts to refine the reviewer assignment process to enhance the integrity of academic peer review. As such, in this paper, we propose an automated pipeline for reviewer-paper matching designed to enhance the accuracy, security, and fairness of reviewer assignments. Our approach integrates advanced natural language processing (NLP) techniques, keyword extraction methods performed by LLMs, and dual similarity measures, specifically Jaccard similarity and

cosine similarity using TF-IDF vectors, to assess the alignment between reviewers and authors' manuscripts. We transform textual data into numerical representations that facilitate precise similarity computations. The evaluation we have carried out using metrics like Mean Reciprocal Rank (MRR), Precision at k (P@k), and Mean Average Precision (MAP), demonstrates that our proposed system significantly outperforms traditional methods.

2 Reviewer Assignment Pipeline

This study implements a four-stage systematic pipeline for automated reviewer-paper matching, integrating NLP, keyword extraction, and similarity measures to enable a scalable and efficient review assignment process.

2.1 Data Loading and Parsing of LLM Outputs

The initial phase of the process focuses on preparing and parsing the dataset, composed of both reviewer profiles (as a list of most cited works with title and abstract for each reviewer) and academic paper descriptions, stored in structured JSON files.

We employed LLAMA 3.2 via the Ollama API [1] to extract keywords from titles, summaries of abstracts, and full abstracts. We adopted the following prompt for authors' papers :

> *"Generate a set of keywords that would help find a reviewer for the following paper* `text_type` *: '*`{text}`*'."*

Similarly, for reviewer publications, we used the prompt :

> *"Propose a set of keywords that define the expertise of the author who wrote a paper with the following* `text_type` *: '*`{text}`*'."*

In these templates, the placeholder `{text}` is dynamically replaced with either the title or abstract or its summary, depending on the specific dataset component being processed. The placeholder `text_type` is replaced with the typology of text (*title, summary, abstract*) that was being used to extract the keywords from. For example, for a paper title, we used the following prompt :*"Generate a set of keywords that would help find a reviewer for the following paper title : 'Entity Typing and Linking using SPARQL Patterns and DBpedia'"*. We utilized the *ollama.chat*() function to interact with the model, formatting the prompts as part of the user messages. LLAMA 3.2 produced lists of keywords tailored to the content of the input text. The outputs included contextually relevant keywords for each author's paper description or reviewer's profile. These keywords aligned the expertise of reviewers with the topics covered in the submitted authors' papers. The entire process was implemented in Python.

2.2 Numerical Representation of Keywords

Keywords are transformed into numerical vectors to enable computational analysis. Specifically, TF-IDF vectorization is employed, which assigns weights to keywords

1. `https://ollama.com/library/llama3.2`

based on their importance within the dataset. This approach ensures that commonly occurring but less informative terms are down-weighted, enhancing the accuracy of similarity assessments.

2.3 Semantic Triple Extraction (OpenIE)

Before constructing the knowledge graph and executing the reviewer assignment, we implemented a preprocessing phase that transforms raw textual content into structured, semantically enriched representations by using OpenIE [2], an unsupervised technique to convert unstructured text into structured semantic triples. It transforms each abstract into a set of (subject, relation, object) triples, providing a foundational structured representation upon which further analysis can be performed.

2.4 CSO Classifier

CSO Classifier is an unsupervised tool that organizes the extracted propositions into a comprehensive taxonomy of computer science topics (Salatino et al. (2019)). By mapping the identified entities from the output of the Open IE and their relationships to the CSO's hierarchical structure, we obtain an enriched semantic representation and we lay the groundwork for constructing a more detailed knowledge graph.

2.5 Knowledge Graph Construction

To build the knowledge graphs from the previously extracted and classified information we proceed as follows. Each node in the graph corresponds to a CSO classifier-defined topic, while edges represent relationships derived from OpenIE ("relation"). To facilitate quantitative analysis and comparison, the knowledge graphs are embedded into vector spaces using the Node2Vec [3] algorithm. The final graph embeddings capture essential structural and semantic features, enabling further analytical tasks.

2.6 Similarity Computation

Two complementary similarity measures are used to assess the match between reviewers and papers : Jaccard Similarity compares the overlap between keyword sets by calculating the intersection over the union of the sets. Cosine Similarity with TF-IDF vectors calculates the angle between the TF-IDF vectors, assessing the alignment of keywords in a high-dimensional space, considering their importance and frequency. The average of the Jaccard and cosine similarity value is computed, offering a balanced metric incorporating direct keyword overlap and nuanced relationships based on term importance. This hybrid approach improves the overall matching process.

2. `https://nlp.stanford.edu/software/openie.html`
3. `https://snap.stanford.edu/node2vec/`

2.7 Reviewer-Paper Matching

The similarity scores from Jaccard and cosine similarity measures are used to match reviewers with papers. The algorithm assigns reviewers to papers based on the highest match, ensuring that each paper is reviewed by experts with closely related interests and expertise. This automated process optimizes relevance in the reviewer-paper matching.

3 Evaluation

This section outlines the evaluation methodology used to assess the performance of the reviewer-paper matching system, including the metrics for effectiveness and the experiments conducted to validate the approach.

3.1 Dataset

The dataset for this study consists of two parts : author papers and reviewer publications. The authors' dataset includes titles and abstracts from 663 papers across domains like AI, databases, and NLP, collected from the EasyChair platform [4] with the authorization of conference organizers. The reviewers' dataset, obtained via Semantic Scholar's API [5], features the 20 most-cited articles for each of the 524 reviewers linked to these papers. All data were anonymized for confidentiality. To enable diverse representation for analysis, papers are represented through their titles, abstracts, or LLM-generated summaries of abstracts. The summarization process utilizes the T5 (Text-to-Text Transfer Transformer) [6] base model, a versatile pre-trained transformer renowned for its state-of-the-art performance in NLP tasks. Fine-tuned on the SciTLDR [7] dataset—specifically curated for summarization in the scientific domain—the model generates summaries capped at 150 tokens to balance informativeness and brevity.

3.2 Evaluation Metrics

To quantitatively evaluate the method, standard information retrieval metrics used in recommender systems are applied. In this context, a reviewer is deemed relevant if they were originally assigned to review the paper in the dataset.

Mean Reciprocal Rank (MRR) : The Mean Reciprocal Rank (MRR) [8] measures the effectiveness of our system in ranking relevant reviewers at the top of the recommendation list. For each paper, the system generates a ranked list of reviewers based on similarity scores calculated from their past work and the candidate paper's content. Higher similarity scores indicate greater topical relevance. The MRR is computed by taking the average of the reciprocal ranks of the first relevant reviewer in each paper's

4. https://www.easychair.org/
5. https://www.semanticscholar.org/
6. https://huggingface.co/docs/transformers/en/model_doc/t5
7. https://huggingface.co/datasets/allenai/scitldr?row=0
8. https://www.evidentlyai.com/ranking-metrics/mean-reciprocal-rank-mrr

list :

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i}$$

where Q is the total number of papers, and rank_i is the rank position of the first relevant reviewer for the i-th paper. MRR values range between 0 and 1, with values closer to 1 indicating that relevant reviewers are consistently ranked higher (Ali et al. (2021)), thus demonstrating better system performance.

Precision at k (P@k) : Precision at k ($P@k$) evaluates the proportion ofthe relevant reviewers among the top k recommendations provided by the system for each paper(Khan et al. (2023)). In this study, we set $k = 3$, aligning with the typical number of reviewers assigned to a manuscript in academic conferences and journals. $P@k$ is defined as :

$$P@k = \frac{Number\ of\ relevant\ reviewers\ in\ top\ k}{k}$$

A higher $P@3$ value, closer to 1, indicates that the top three suggested reviewers are mostly relevant, demonstrating the system's effectiveness in matching reviewers to papers. However, $P@k$ focuses only on the top k reviewers and may miss relevant reviewers ranked lower.

Mean Average Precision (MAP) : The Mean Average Precision (MAP) [9] provides an overall precision score by averaging the precision across all relevant reviewers for each paper and then averaging over all papers. For each paper, it calculates the precision at each rank position where a relevant reviewer is found and then computes the mean of these precision values. MAP is calculated as(Zhang et al. (2023)) :

$$\text{MAP} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{R_i} \sum_{r=1}^{R_i} \text{Precision}(r) \right)$$

where N is the total number of papers, Ri is the number of relevant reviewers for the $i-th$ paper, and Precision(r) is the precision at rank r. Unlike P@k, MAP considers all relevant reviewers throughout the entire ranked list, not just the top-k positions, and thus may yield a higher value. MAP is beneficial for assessing how consistently the system ranks relevant reviewers across different papers, including those ranked beyond the top positions.

3.3 Experimental Setup

We evaluated our reviewer assignment system by comparing its assignments to actual reviewer allocations from EasyChair. Experiments were conducted using different textual inputs to assess their impact on similarity computation between papers and reviewers. The variations tested included :

— paper titles ;
— full abstracts ;
— summarized abstracts.

9. https://www.evidentlyai.com/ranking-metrics/mean-average-precision-map

The goal was to identify which type of textual content most effectively captures a paper's core topics to improve reviewer matching accuracy.

4 Results

In this section, we present the results of our experiments on reviewer-paper matching using cosine similarity, Jaccard similarity, and their average across different text fields (titles, summaries, and abstracts).

Table 1 shows results for summary-based matching, where cosine similarity achieves the highest Mean Reciprocal Rank (MRR of 0.6534) and Mean Average Precision (MAP of 0.2522). This indicates that cosine is particularly effective in aligning keywords for summaries, achieving both strong top-ranked matches and overall match quality. As a baseline for comparison, we utilized our previous work (*anonymized*), which did not incorporate the use of LLMs and instead employed traditional natural language processing techniques such as tokenization, stemming, and TF-IDF applied directly to the titles, summaries, and abstracts of papers to represent reviewer and paper profiles. In contrast, the current study applies these preprocessing techniques to the outputs generated by LLMs, enhancing the semantic representation and matching accuracy over our prior approach.

	LLM Method			Baseline (no LLM)		
	Cosine Similarity	Jaccard Similarity	Cosine-Jaccard Average Similarity	Cosine Similarity	Jaccard Similarity	Dot Product Similarity
MRR	**0.6534**	0.6501	0.6503	0.238	0.3095	0.5476
P@K	**0.3674**	0.3557	0.3550	0.0845	0.1673	0.3202
MAP	0.2522	0.2462	0.2415	0.1369	0.1964	**0.3452**

TAB. 1 – *Reviewer matching using the summaries of abstracts.*

Table 2 focuses on titles, with Jaccard achieving the highest MRR and MAP, highlighting that exact term overlap works well for concise text. Cosine also performs well, though slightly less than Jaccard.

	LLM Method			Baseline (no LLM)		
	Cosine Similarity	Jaccard Similarity	Cosine-Jaccard Average Similarity	Cosine Similarity	Jaccard Similarity	Dot Product Similarity
MRR	0.6715	**0.6848**	0.6401	0.3669	0.0769	0.6111
P@K	0.3627	**0.3645**	0.3613	0.1409	0.0192	0.1809
MAP	0.2570	**0.2621**	0.2440	0.2273	0.0384	0.3611

TAB. 2 – *Reviewer matching using the titles of papers.*

Table 3 presents abstract-based results, where the results with Jaccard excel, especially in MRR and MAP, likely due to its ability to handle the complexity of detailed text. The cosine and the average similarities show weaker performance in this context. Table 4 presents the evaluation results for the CSO-Classifier relation. The Cosine-Jaccard Average Similarity consistently outperforms the other similarity measures,

	LLM Method			Baseline (no LLM)		
	Cosine Similarity	Jaccard Similarity	Cosine-Jaccard Average Similarity	Cosine Similarity	Jaccard Similarity	Dot Product Similarity
MRR	0.6584	**0.7112**	0.6364	0.6958	0.5737	0.6263
P@K	**0.3703**	0.3450	0.3674	0.3259	0.3125	0.3071
MAP	0.2551	0.2643	0.2455	**0.3995**	**0.5881**	0.3737

TAB. 3 – *Reviewer matching using the abstracts of papers.*

particularly in MRR and P@3. Knowledge graph-based matching produces the highest

	Cosine Similarity	Jaccard Similarity	Cosine-Jaccard Average Similarity
MRR	0.7619	0.7424	**0.7857**
P@3	0.3445	0.3471	**0.3920**
MAP	0.2539	0.2475	**0.2936**

TAB. 4 – *Reviewer matching using CSO-Classifier with relation edge*

overall relevance scores across MRR, MAP, and P@3, particularly when employing the cosine-Jaccard average similarity measure. In contrast, abstract-based matching exhibits comparatively lower relevance on MRR and MAP, especially under the Jaccard similarity measure, when compared to knowledge graph-based results. Titles, due to their brevity and focused content, benefit from Jaccard similarity but nonetheless fail to outperform abstracts or knowledge graph-based methods in terms of MRR and MAP. Furthermore, the combined cosine-Jaccard similarity consistently improves reviewer assignment performance, especially within knowledge graph-based approaches.

5 Conclusions

In this paper, we address the challenge of efficiently assigning reviewers to papers to ensure quality reviews. Our study utilized a dataset comprising 663 papers from 85 conferences and 524 reviewer profiles. For each reviewer, we analyzed their 20 most-cited articles. To represent the content of authors' and reviewers' papers, we explored four strategies : using only titles, only abstracts, summaries of abstracts and a knowledge graph-based approach. In the first three strategies, we employed an LLM to extract keywords and applied TF-IDF vectorization to assign weights to the extracted keywords. Additionally, we incorporated a knowledge graph-based method for reviewer assignment, constructing the graphs through an Open Information Extraction pipeline and deriving thematic concepts using the CSO classifier. We then measured the similarity between authors' and reviewers' papers using three metrics : Jaccard similarity, cosine similarity, and their average. Our experimental results demonstrate that the proposed method outperforms our previous framework across all evaluation metrics : MRR, P@k, and MAP. Future directions for this research include exploring

more sophisticated techniques for text representation, such as embeddings generated by transformer-based models like BERT or GPT. Additionally, testing the scalability and generalizability of our method on larger and more diverse datasets will be crucial for its adoption in real-world conference management systems.

Références

Abduljaleel, A. Q. et al. (2021). Reviewer assignment using weighted matching and hungarian algorithm. *Turkish Journal of Computer and Mathematics Education (TURCOMAT) 12*(7), 619–627.

Ali, Z., I. Ullah, A. Khan, A. Ullah Jan, et K. Muhammad (2021). An overview and evaluation of citation recommendation models. *Scientometrics 126*, 4083–4119.

Arabzadeh, N., S. Ebrahimi, S. Salamat, M. Bashari, et E. Bagheri (2024). Reviewerly : Modeling the reviewer assignment task as an information retrieval problem. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 5554–5555.

Hoang, D. T., N. T. Nguyen, B. Collins, et D. Hwang (2021). Decision support system for solving reviewer assignment problem. *Cybernetics and Systems 52*(5), 379–397.

Jin, Y., Q. Zhao, Y. Wang, H. Chen, K. Zhu, Y. Xiao, et J. Wang (2024). Agentreview : Exploring peer review dynamics with llm agents. *arXiv preprint arXiv :2406.12708*.

Khan, F., M. Al Rawajbeh, L. K. Ramasamy, et S. Lim (2023). Context-aware and click session-based graph pattern mining with recommendations for smart ems through ai. *IEEE Access*.

Latypova, V. (2023). Reviewer assignment decision support in an academic journal based on multicriteria assessment and text mining. In *2023 IX International Conference on Information Technology and Nanotechnology (ITNT)*, pp. 1–4. IEEE.

Payan, J. et Y. Zick (2021). I will have order ! optimizing orders for fair reviewer assignment. *arXiv preprint arXiv :2108.02126*.

Salatino, A. A., F. Osborne, T. Thanapalasingam, et E. Motta (2019). The cso classifier : Ontology-driven detection of research topics in scholarly articles. In *Digital Libraries for Open Knowledge : 23rd International Conference on Theory and Practice of Digital Libraries, TPDL 2019, Oslo, Norway, September 9-12, 2019, Proceedings 23*, pp. 296–311. Springer.

Stelmakh, I., J. Wieting, G. Neubig, et N. B. Shah (2023). A gold standard dataset for the reviewer assignment problem. *arXiv preprint arXiv :2303.16750*.

Tan, S., Z. Duan, S. Zhao, J. Chen, et Y. Zhang (2021). Improved reviewer assignment based on both word and semantic features. *Information Retrieval Journal 24*, 175–204.

Tyser, K., B. Segev, G. Longhitano, X.-Y. Zhang, Z. Meeks, J. Lee, U. Garg, N. Belsten, A. Shporer, M. Udell, et al. (2024). Ai-driven review systems : Evaluating llms in scalable and bias-aware academic reviews. *arXiv preprint arXiv :2408.10365*.

Xu, L., D. Zeng, J. Dai, et L. Gui (2022). Combining coverage with tmps for reviewer assignment. In *2022 International Conference on Intelligent Education and Intelligent Research (IEIR)*, pp. 107–113. IEEE.

Zhang, T., Y. Zhang, M. Xin, J. Liao, et Q. Xie (2023). A light-weight network for small insulator and defect detection using uav imaging based on improved yolov5. *Sensors 23*(11), 5249.

Zhao, X. et Y. Zhang (2022). Reviewer assignment algorithms for peer review automation : A survey. *Information Processing & Management 59*(5), 103028.

Summary

Dans cet article, nous explorons l'utilisation des grands modèles de langage (LLM) pour l'organisation de conférences, en particulier pour l'affectation des réviseurs en fonction du contenu des soumissions. Assigner des réviseurs pertinents aux articles de recherche est une étape cruciale pour garantir la qualité des conférences et des ateliers scientifiques. Pour mener cette étude, nous avons constitué un ensemble de données comprenant 663 articles issus de 85 conférences et 524 profils de réviseurs, principalement dans les domaines du Web sémantique et de l'informatique. À l'aide d'un grand modèle de langage, des mots-clés ont été extraits pour chaque auteur et chaque réviseur. Nous avons ensuite expérimenté différentes mesures de similarité et stratégies de représentation de ces mots-clés afin d'évaluer la pertinence des réviseurs pour chaque article. Les résultats expérimentaux confirment l'efficacité de notre approche pour améliorer le processus d'assignation des réviseurs aux articles.