# Retrieval Augmented Generation on Large Technical Reports

Henri Scaffidi[1]

Melinda Hodkiewicz[2], Caitlin Woods[1]
1. Computer Science and Software Engineering
2. Mechanical Engineering
University of Western Australia

Nicole Roocke
CEED Client: MRIWA

**Abstract**

*The Minerals Research Institute of Western Australia (MRIWA) has a collection of more than 300 technical reports, each containing research data and insights of value to today's minerals industry. However, the reports lack a standardized format and traditional search techniques (using keywords) are inefficient and ineffective for information extraction. Retrieval Augmented Generation (RAG) enables Large Language Models (LLMs), a generative artificial intelligence, to ask and answer questions of the contents in, and across, reports. This project develops a RAG pipeline, exploring LLM capabilities to unlock MRIWA report insights and data. The approach utilises LLMs to 1) construct Knowledge Graphs (KGs) from MRIWA's technical reports and 2) answer natural language queries related to the reports. The GraphRAG method is compared using four distinct KG schemas and asked several competency questions. Answers contain true statements but miss key details contained within model responses. A MRIWA-specific generalised KG schema of five distinct abstract concepts improves the GraphRAG responses by increasing the quantity of KG entities by 30%, compared to using a MRIWA-specific expanded KG schema of eight concepts. Specifying domain-specific KG schemas appears to improve GraphRAG's answers about minerals industry text.*

## 1. Introduction

The Minerals Research Institute of Western Australia (MRIWA) has a collection of more than 300 technical reports, often exceeding 150 pages each. These reports, compiled over 40 years, contain data and insights of potential value to Western Australia's (WA) minerals industry and the research community. Content ranges from table-, graph-, and text-based data. The primary challenge is efficiently extracting value from this long and technical content. Currently, the public-facing system relies on keyword search on project summaries to identify potentially relevant reports. This process can fail due to differences between full reports and their respective summaries. Additionally, the process of locating specific information in the full reports is both cumbersome and inefficient. Aggregating information across reports requires specialized expertise.

The public release of Large Language Models (LLMs), such as OpenAI's Generative Pre-trained Transformer (GPT), has seen many organisations seek to leverage the power of this generative artificial intelligence (AI), often yielding faster task completion and higher quality solutions (Dell'Acqua et al., 2023). However, LLMs are typically limited to their pre-trained data, causing the AI to generate false content as if it were true (a phenomenon known as hallucination) when additional knowledge is required.

Retrieval Augmented Generation (RAG) allows the LLM to access and incorporate external information at query time. With RAG, LLM responses can be grounded in domain-specific sources, countering hallucination and increasing verifiability.

Baseline RAG takes the user's query, compares it to a set of text chunks from documents, and retrieves the most semantically similar text chunks (a process known as vector similarity search). The retrieved text chunks are provided to the LLM to assist in answering the query. This method was implemented in MRIWA's past RAG trial. The approach is scalable and efficient but falters when 1) the query requires information to be aggregated across documents (aggregation queries) and 2) the query requires the system to derive insights not explicitly mentioned in reports (complex semantic queries). In contrast, Knowledge Graphs (KGs) illustrate concept relatedness and provide a holistic view of the documents (Buehler, 2024), potentially countering the baseline RAG weaknesses, and therefore motivating the recent surge in KG-RAG research.

This research project aims to develop a RAG pipeline for MRIWA, their stakeholders in industry, and the research community. The approach utilises LLMs to 1) construct KGs from MRIWA's technical reports and 2) answer natural language queries related to the reports. This approach seeks to unlock the value embedded in these historical documents, integrate it with current information, and foster new insights in the WA minerals industry.

# 2. Process

## 2.1 Schema Development

The KG schema is a set of concepts which defines the types of entities in the KG. The schema, therefore, dictates the type of information to be extracted from MRIWA report text during KG construction. For example, if "Organisation" is a concept defined in the KG schema, "MRIWA" would be an entity extracted from the text. This extracted content should be unique and pertinent to the report it is sourced from, else there is a risk of introducing noise into the KG, which may hinder information retrieval.

Through manual annotation of report content, automated Named Entity Recognition trials, and analysis of the Common Core Ontologies (Jensen et al., 2024) and the Critical Minerals Ontology (Davarpanah et al., 2024), two candidate schemas for KG construction are derived (see Table 1): *Generalised Schema* and *Expanded Schema.*

## 2.2 GraphRAG

Microsoft's GraphRAG approach is designed to answer complex semantic queries and aggregation queries, being two weaknesses of MRIWA's baseline RAG trial (Edge et al., 2024). This approach is utilized in the current project. As shown in Figure 1, the methodology consists of two phases: construction and query.

| **Generalised Knowledge Graph Schema** | |
|---|---|
| ***Entity Type*** | ***Description*** |
| Object | All objects. |
| Naturally_Occurring_Object | Objects formed naturally. |
| Processed_Object | Objects formed through human intervention. |
| Process | All processes/procedures. |
| Site_Location_Boundary | Immaterial spatial regions. |
| Organisation | Person/s organised together for a purpose. |
| **Expanded Knowledge Graph Schema** | |
| ***Entity Type*** | ***Description*** |
| Object | All objects. |
| Naturally_Occurring_Object | Objects formed naturally. |
| Processed_Material | Objects formed through human intervention that are used during a process. |
| Manufactured_Product | Products of a manufacturing process. |
| Process | All processes/procedures. |
| Natural_Process | Naturally occurring processes. |
| Lab_Process | Small-scale lab-based processes. |
| Industrial_Process | Large-scale industrial processes. |
| Site_Location_Boundary | Immaterial spatial regions. |
| Organisation | Person/s organised together for a purpose. |

**Table 1** Generalised Knowledge Graph Schema and Expanded Knowledge Graph Schema.

Construction involves GPT-4o-Mini processing seven of MRIWA's technical reports. Entities and relations are extracted from each 300-token chunk of the reports based on the KG schema. Entities and relations with identical names are then grouped and summarised, forming the KG. The Leiden algorithm (Traag et al., 2019) is used to identify semantically related communities within the KG, which are further summarised into community reports.

The query phase first vectorises the user's prompt to identify the top 20 semantically related entities in the KG. These entities, and other closely connected entities, relations, text chunks, and community reports, are retrieved. These data sources are ranked according to their effectiveness in addressing the user's query, and finally passed to GPT to provide a knowledge-grounded response.

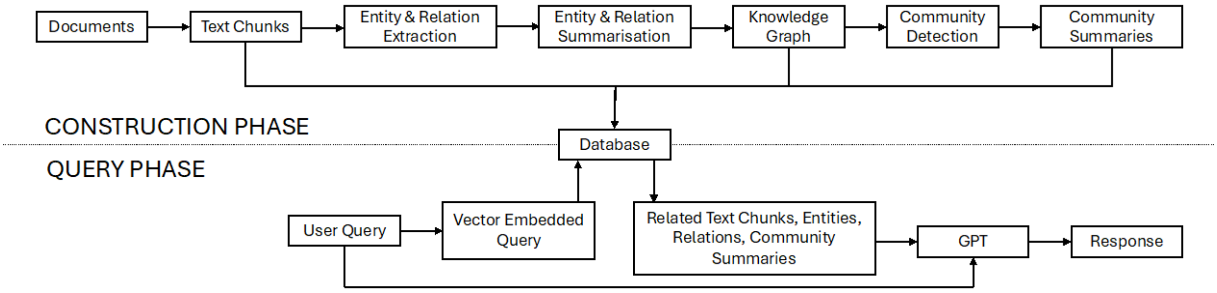

**Figure 1** Pipeline for knowledge graph construction and question answering.

# 3. Results and Discussion

GraphRAG is trialed using four distinct KG schemas for KG construction: Generalised KG Schema (**GS**), Expanded KG Schema (**ES**), Auto-Generated KG Schema (**AS**), and Schema-less (**SL**). The AS is yielded through the GraphRAG "Prompt Tuner" which uses GPT to derive the following set of entity types: *{chemical process, mineral, geological survey, geochemistry, exploration technique, sample, formation, project, research study}.* The SL pipeline leaves the KG schema undefined, and instead prompts GPT to "identify all entities needed from the text in order to capture the information and ideas in the text".

## 3.1 Competency Questions

Once the construction phase is complete, each pipeline is tested on 15 competency questions provided by MRIWA. These questions include keyword search, summarisation, and complex semantic queries, identifying CSIRO's project involvement in specific capacities, for example. Each response is rated from 0 to 3 based on the quantity of incorrect statements and the quantity of model answer information present. This scoring system is a preliminary metric and is to be refined throughout the remaining course of the project. The results of this analysis are displayed in Table 2.

| | **Generalised KG Schema** | **Expanded KG Schema** | **Auto-Generated KG Schema** | **Schema-less** |
|---|---|---|---|---|
| Mean | **1.10** | 0.97 | 0.87 | 0.87 |
| SD | 0.80 | 0.67 | 0.78 | 0.78 |

**Table 2** Results for competency question answers.

Responses comprise of multiple paragraphs which contain references to retrieved sources. On average, responses are rated 0.95, as most contain less than half the model answer facts. The few-shot prompts used during construction and querying are not tailored to the minerals research domain and may contribute to the low scores.

## 3.2 KG Schema Performance

The GS pipeline scores approximately 13% higher than ES and 26% higher than both AS and SL. The KG schema specifies to the LLM which entity types should be extracted from MRIWA's reports. AS and SL, effectively, permit the LLM to decide what types of entities to extract. This causes less valuable information to be extracted, summarised, and later retrieved, likely resulting in the decrement of response quality compared to GS and ES**.**

To understand why GS outperforms ES, a comparison of the quantity of entities, relations, and communities within the constructed KGs is conducted (see Table 3). The GS pipeline extracts approximately 30% more entities than ES. This is likely due to GS specifying fewer concepts at a higher level of abstraction, meaning the concepts are more distinct, and thus clearer for the non-expert LLM to apply during entity extraction. This finding may vary in non-technical domains where the LLM could better understand the concepts in an expanded KG schema.

| | GS | ES | AS | SL |
|---|---|---|---|---|
| # Entities | 183015 | 139626 | 140916 | 128718 |
| # Relations | 40220 | 38634 | 41164 | 43974 |
| # Communities | 4029 | 3706 | 3817 | 3989 |

**Table 3** Results for knowledge graph analysis.

Since the GS pipeline extracts more entities, the KG constructed has a higher likelihood of containing facts necessary to answer a query. For example, the AS pipeline tags 42 instances of CSIRO, whereas GS**,** ES**,** and SL tag approximately 60 each. Effectively, this means 18 facts about CSIRO are missed during KG construction within the AS pipeline. The effect of this is evident when the pipelines are asked "Which MRIWA reports has CSIRO been involved with as a researcher?", a question which AS scores 0.5 for, and GS**,** ES**,** and SL score 2.0, 1.5, and 2.0 respectively.

Given that GS and ES tag comparable quantities of CSIRO instances, similar performance is expected on the aforementioned CSIRO question. Yet, ES performs worse. In this case, this is due to the summarised CSIRO entity in the ES KG missing one key fact needed to answer this competency question. Whilst not definitive, this is likely the result of using a non-deterministic LLM, where given the same set of CSIRO instances, LLM-summarisation of CSIRO may not be identical. This issue may be solved by conducting entity extraction multiple times to detect additional entities that the LLM may have missed (Edge et al., 2024).

# 4. Conclusions and Future Work

MRIWA's technical reports contain untapped value and demand a solution to make information extraction more efficient. GraphRAG extracts value from MRIWA's reports, models the content as a KG, and enables democratised access to the information via natural language queries. Generally, the responses make true statements but miss key information needed to answer competency questions.

The KG schema used during KG construction is shown to impact the quality of RAG's responses, where a generalised KG schema of five distinct and abstract concepts leads to higher response scores. The generalised KG schema results in more entities being extracted from MRIWA's reports, increasing the likelihood that necessary facts are contained within the constructed KG. In addition to the KG schema, GraphRAG may be further enhanced by including domain-specific examples of entity extraction within the LLM prompts used during KG construction. This will be tested if time permits.

The non-deterministic nature of LLMs may result in omission of facts during entity extraction, therefore future work could investigate countering this issue through multiple gleanings. In addition, future research could explore whether the KG schema impacts GraphRAG in non-technical domains, where an LLM may better understand the entity types and thus could adequately apply an expanded KG schema.

## 5. Acknowledgements

This project was made possible thanks to the support, mentorship, and dedication of my academic supervisors, Melinda Hodkiewicz and Caitlin Woods, as well as my CEED mentor from MRIWA, Nicole Roocke. I am grateful to MRIWA, Jeremy Leggoe, and Kimberlie Hancock for providing this valuable research opportunity.

## 6. References

Buehler, M. J. (2024). Generative retrieval-augmented ontologic graph and multiagent strategies for interpretive large language model-based materials design. ACS Engineering Au, 4(2), 241-277.

Davarpanah, A., Babaie, H. A., & Elliott, W. C. (2024). Knowledge-based query system for the critical minerals. Applied Computing and Geosciences, 22, 100167.

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., ... & Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. Harvard Business School Technology & Operations Mgt. Unit Working Paper, (24-013).

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., ... & Larson, J. (2024). From local to global: A graph rag approach to query-focused summarization. arXiv preprint arXiv:2404.16130.

Jensen, M., De Colle, G., Kindya, S., More, C., Cox, A. P., & Beverley, J. (2024). The Common Core Ontologies. arXiv preprint arXiv:2404.17758.

Traag, V. A., Waltman, L., & Van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. Scientific reports, 9(1), 1-12.