Submission code: JAWL6GBX40

See this and other abstracts on our website: https://home.sbgf.org.br/Pages/resumos.php

From Specialist Audio to Knowledge Graphs: An LLM Approach for Geophysical Image Metadata Generation

Pablo Machado Barros (Cenpes-Petrobras), Victor do Nascimento Gomes (Universidade Federal do Rio Grande do Norte), Adriane Gomes Pinheiro Praxedes (Universidade Federal do Rio Grande do Norte), Quelita Miriam (Universidade Federal do Rio Grande do Norte), Poliana Araujo (Universidade Federal do Rio Grande do Norte), Ivanovitch Medeiros Dantas Da Silva (Universidade Federal do Rio Grande do Norte), Luiz Affonso Henderson Guedes De Oliveira (Universidade Federal do Rio Grande do Norte), Cristhian Alberto Celestino Cortez (Cenpes-Petrobras), Tiago Ilipronti Girardi (Cenpes-Petrobras), Tiago Barros (Universidade Federal do Rio Grande do Norte), Luis Fernando Mendes Cury (Cenpes-Petrobras), Leandro da Silva Sadala Valente (Cenpes-Petrobras), Lucas Gondim Miranda (Cenpes-Petrobras)

From Specialist Audio to Knowledge Graphs: An LLM Approach for Geophysical Image Metadata Generation

Abstract Summary

Interpreting geophysical and geological images requires specialized expertise, yet capturing this knowledge systematically for broader use is challenging. Current methods for generating descriptive metadata are often manual and resource-intensive, hindering efficient data discovery and sharing. This paper proposes a novel methodology leveraging specialist audio descriptions of images. The workflow begins with expert audio insights, which are then transcribed using Large Language Models (LLMs). Named Entity Recognition (NER) techniques extract key geological features and parameters, creating a knowledge graph (KG) that models relationships between images and concepts. This approach aims to automate rich metadata generation and uncover new insights, especially in petroleum exploration.

Introduction

Interpreting geophysical and geological data, such as seismic images, subsurface maps, and well logs, is critical for resource exploration, hazard assessment, and environmental monitoring applications. However, meaningful insights from these complex datasets often rely on domain specialists' implicit, experience-based knowledge. This includes recognizing subtle patterns, understanding geological contexts, and being familiar with specific acquisition and processing techniques. Such expert knowledge is difficult to formalize and scale using traditional data management systems, often remaining siloed in individual interpretations or dispersed across project documents, resulting in inefficiencies and knowledge loss over time.

Current methods for managing geophysical image data typically depend on simple file naming conventions or limited, manually curated metadata. Although databases exist, generating consistent and comprehensive metadata that accurately reflects expert interpretation is labor-intensive, costly, and complex. This lack of structured, searchable knowledge hinders effective querying of large image repositories based on geological features, interpretative insights, or processing techniques. For example, retrieving all seismic sections that show a specific fault type or were processed with a particular algorithm is still challenging.

Recent advancements in Artificial Intelligence (AI), especially in Natural Language Processing (NLP) with Large Language Models (LLMs) Minaee et al. (2025), Named Entity Recognition (NER) Li et al. (2022), and Knowledge Graphs (KGs) Peng et al. (2023) offer promising solutions to address these challenges. LLMs can interpret human language at scale, NER can extract key terms and concepts from text, and KGs can model complex relationships between entities. This paper presents a methodology to transform expert audio descriptions of geophysical images into structured, queryable knowledge. The workflow includes capturing audio commentary, transcribing it with LLMs, extracting relevant entities via NER, and building a domain-specific knowledge graph. The goal is to automate rich metadata generation, enabling advanced data exploration and knowledge discovery, particularly for petroleum geology applications.

Method and or Theory

The proposed methodology integrates several AI techniques to transform unstructured specialist knowledge, captured via audio, into a structured knowledge graph representation suitable for querying and analysis.

Figure 1 illustrates the overall workflow adopted in this study, encompassing audio acquisition, transcription, entity recognition, and graph construction. The process consists of four main steps:

1. **Audio Description Acquisition** involves capturing specialists' interpretations and commentary on specific geophysical images (e.g., seismic sections, well-log displays) as audio recordings. An interactive Streamlit application enables users to upload each image, record an audio description in Portuguese, and securely store the original file and its metadata for downstream processing.

2. **LLM Transcription** utilizes Gemini 1.5 to convert the recorded audio directly into text. Gemini 1.5 was selected for its demonstrated robustness in handling technical vocabulary and varied accents typical of geophysical discourse.

3. **Named Entity Recognition (NER)** applies GPT-4-o-mini, guided by prompt engineering with few-shot examples and Chain-of-Thought (CoT) reasoning Chen et al. (2024). The set of entity types such as fault styles, stratigraphic units, seismic attributes, processing techniques, and interpretative actions, was defined based on the domain-specific taxonomy presented by Freitas et al. (2023). This structured ontology ensures that extracted entities align with geoscientific concepts of relevance.

4. **Knowledge Graph Construction** involves modeling the extracted entities and their interconnections based solely on shared entities: each *Image* node is linked to every other *Image* node with which it shares at least one extracted entity. During graph population, an edge is created between two image nodes whenever they have one or more entities in common; the edge's weight equals the count of shared entities, and it is annotated with the list of those entity labels.

Figure 2 provides a simplified example of the resulting knowledge graph. It illustrates how semantic connections emerge between images through shared entities. For instance:

- Images 5 and 9 share the entity *elevação do Rio Grande*;
- Images 5 and 16 share *zona de fratura Rio Grande* and *elevação do Rio Grande*;
- Images 53 and 54 share *zona de transição*.

This modeling sometimes called "modeling by label" when using qualitative tags, or "modeling by value" when numerical values are involved enables complex traversals of the graph. One can query for all images with specific geological features or compare clusters of pictures connected by common concepts, facilitating deeper, relationship-aware knowledge retrieval beyond simple keyword search.

Figure 2 illustrates an example output of the graph construction step. This visualization highlights how geoscientific entities are semantic bridges between otherwise unconnected images.

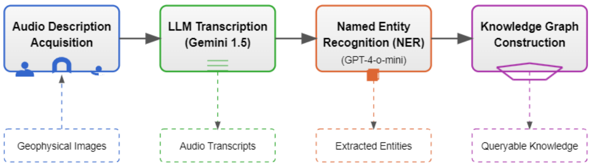

Figure 1: Overview of the workflow: from audio description acquisition to knowledge graph construction.

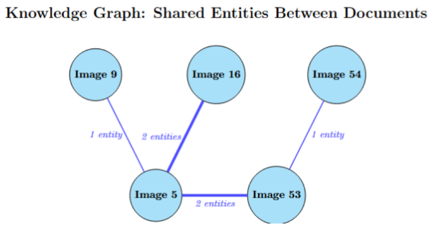

Figure 2: Simplified knowledge graph showing connections between documents based on shared entities. The edge thickness reflects the number of entities in common.

Results

The proposed methodology was applied to audio descriptions of geophysical images from two scientific studies, narrated by specialists in Portuguese. After transcription with Gemini 1.5 and entity extraction via GPT-4-o-mini, the corpus revealed frequent mentions of key geological structures, such as do Rio Grande Rise and the Borborema Province (Figure 3).

A word cloud was generated using entities categorized under ESTRUTURA FISICA to highlight frequently mentioned physical structures (Figure 4). A knowledge graph constructed from the extracted entities showed two main clusters reflecting the original studies, with some cross-links indicating shared geological themes. This structure enables not only metadata storage but also advanced semantic queries.

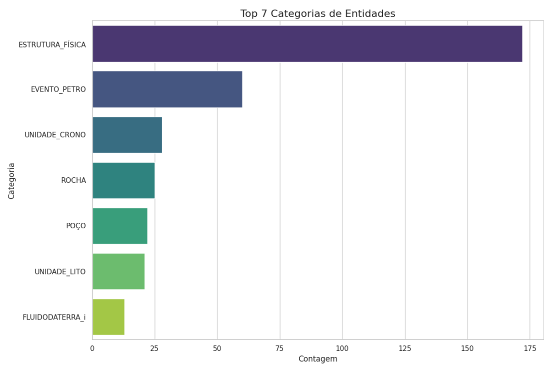

Figure 3: : Word cloud showing the most frequently mentioned physical structures (category: ESTRUTURA FISICA) extracted from expert descriptions.

Figure 4: Distribution of entity categories extracted from specialist descriptions across the corpus, highlighting the predominance of ESTRUTURA FISICA category.

The knowledge graph (KG) (Figure 5) construction revealed an interesting pattern with two main communities in the graph. This clear separation corresponds to the two distinct scientific articles from which the images originated, each with its own terminology and geological focus.

Despite this separation, the graph revealed interesting connections between images from different articles, indicated by edges crossing between communities. These connections represent shared geological concepts, suggesting that certain structural features are consistently identified across independent studies.

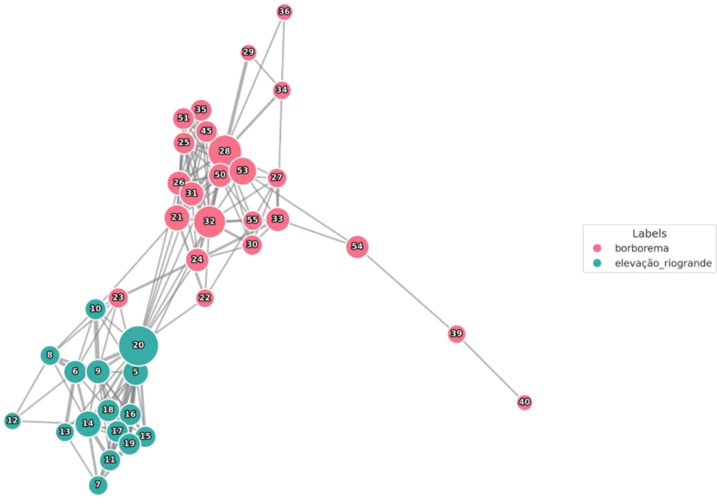

Figure 5: Knowledge graph showing relationships between images: pink nodes for the Borborema Province article and green nodes for the Rio Grande Rise article.

Conclusions

The proposed methodology showcases the potential of LLMs and knowledge graphs to transform expert geophysical audio descriptions into structured, searchable metadata. By automating entity extraction and relationship modeling, the workflow enhances knowledge retrieval and enables advanced graph-based analyses such as centrality and community detection. The implemented application streamlines metadata generation and supports more efficient geoscientific knowledge sharing. Future directions include refining relationship types, integrating with existing platforms, and exploring retrieval-augmented approaches to support deeper semantic querying and hypothesis generation in geological datasets.

References

Chen, Y., Y. Xie, Z. Jiang, and X. Ren, 2024, A systematic survey of prompt engineering techniques: arXiv preprint arXiv:2406.06608.

Freitas, C., E. de Souza, M. C. Castro, T. Cavalcanti, P. F. d. Silva, and F. C. Cordeiro, 2023, Recursos linguísticos para o PLN específico de domínio: o petrolês: Linguamática, **15**, 51–68.

Li, J., A. Sun, J. Han, and C. Li, 2022, A survey on deep learning for named entity recognition: IEEE Transactions on Knowledge and Data Engineering, **34**, 50–70.

Minaee, S., T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, and J. Gao, 2025, Large language models: A survey: arXiv preprint arXiv:2402.06196.

Peng, C., F. Xia, M. Naseriparsa, and F. Osborne, 2023, Knowledge graphs: Opportunities and challenges: Artificial Intelligence Review, **56**, 13071–13102.