

AI for Data Mining and Knowledge Extraction for Mineral Exploration in Greenland

Pierpaolo Guarnieri, Nikolai Andrianov, Jakob Molander

GEUS, Geological Survey of Denmark and Greenland

Mineral exploration in Greenland has generated thousands of geological field reports, mostly stored as scanned PDFs that are usually not searchable. To extract useful geological knowledge from this vast archive at scale, automated methods are needed to process dense technical text, tables, figures, and hand-drawn diagrams within each document.

The geological datasets available for mineral exploration in Greenland are typically multimodal, multidimensional, heterogeneous, complex, and uncertain, with missing values and noise. This preliminary project focuses on a selected area in East Greenland, Jameson Land, which represents an ideal test site due to a large geological and geophysical dataset, a long mineral-exploration history, and well-known mineral occurrences. The aim is to explore how generative AI and ML can leverage existing data to recognize areas with high mineral prospectivity, with a particular focus on key resources, and the identification of hidden patterns in the data.

The AI approach will enable data cleaning and extraction of geological indicators through automated tasks for large-scale screening, reorganizing, and extraction of information from unstructured data, such as scanned PDFs, using a search routine followed by LLM analysis of the text files. The first step consists of document parsing, converting each typed PDF report to a Markdown text file, using an OCR that faithfully reconstructs not only running text but also the spatial semantics of tables, the descriptive content of figures, and the symbolic annotation of diagrams and technical drawings, running on a local GPU. The second step involves analyzing the reports for established keywords associated with a specific mineral system and counting hits, generating georeferenced locality names, and finding localities that co-occur with attribute keywords in the same paragraph and producing a list of candidates for further review. The third step follows the LLM confirmation, partitioning parsed documents into chunks that are semantically coherent, contextually self-contained, and optimally sized for downstream embedding. During this step, an LLM reads the relevant text passages and decides whether each recognized locality is genuinely associated with the established keywords.

Preliminary results from reports covering Jameson Land suggest that co-occurrence of locality names with mineral-system keywords provides a tractable first pass screening signal, substantially reducing the manual review burden. The conference contribution will present the implemented workflow, an assessment of retrieval quality on a test document set, and a discussion of failure modes specific to heterogeneous geological report collections. If validated, the approach is directly transferable to other frontier regions with comparable legacy data archives.