SpaceOps-2025, ID # 436

Knowledge Graph and Retrieval Augmented Generation based anomaly interpretation and procedure generation, a Sentinel 2 processing centre demonstrator

Pierre Ribardière[a], Barbara Do Couto Vidal[a], Arnaud Le Carvennec [a]*

[a] *CS Group – Sopra Steria, France*
* Corresponding Author, *arnaud.lecarvennec@cs-soprasteria.com*

Abstract

The integration of advanced AI technologies into mission-critical operations is transforming how space agencies and operators manage complex systems. A key innovation in this domain is the combination of **Knowledge Graphs** and **Retrieval-Augmented Generation** to enhance chatbot functionality. This approach significantly improves the ability of conversational agents to understand, explain, and resolve anomalies in high-stakes environments such as satellite missions.
A demonstrator system developed by CS showcases the practical application of this integration using data from the **Copernicus Sentinel-2** mission, specifically focusing on the **image processing chain**. The solution supports operators by analyzing anomalies and recommending operational procedures, drawing on both structured and unstructured knowledge.
Knowledge Graphs represent relationships between system components, algorithms, mission parameters, and historical anomaly data. This structured representation forms a network of interconnected nodes and edges, enabling the system to map complex interdependencies. In the demonstrator, the Knowledge Graph captures detailed information about the TMI processing chain, enhancing the system's contextual understanding.
Retrieval-Augmented Generation complements this by combining large language models with real-time information retrieval from technical documentation, expert analyses, and the Knowledge Graph itself. RAG enables the chatbot to answer domain-specific queries with up-to-date and relevant insights, going beyond the limitations of static AI models.
By leveraging **shadow prompting** and optimization techniques, the chatbot allows users to input anomaly descriptions and receive:

- **Potential causes**, based on historical cases and component relationships.
- **Diagnostic procedures** generated using the contextual knowledge from the graph and retrieved data.
- **Operational resolution steps**, aligned with prior incident outcomes and system-specific logic.

This integrated approach not only improves the quality of information provided to operators but also accelerates decision-making, enhances traceability, and reduces the risk of error during mission-critical tasks.
While the current demonstrator is tailored to the image processing chain, the architecture and methodology are scalable and adaptable to other mission operations, control centres, and space systems. This technology represents a significant step forward in AI-assisted mission support, laying the groundwork for safer, smarter, and more efficient space operations.

Keywords: Artificial intelligence, Large Language Model, Knowledge Graphs, operation support, Retrieval-Augmented Generation

Acronyms/Abbreviations

Definition	Acronyms/Abbreviations
Knowledge Graphs	KG
Retrieval-Augmented Generation	RAG
Copernicus Sentinel-2	S2
Large Language Model	LLM

1. Introduction

As satellite missions grow in complexity and data volume, the need for intelligent systems that support operational decision-making becomes increasingly critical. Traditional rule-based systems are limited in their ability to adapt to novel scenarios or provide insight across interconnected processes. Recent advancements in Artificial Intelligence (AI), particularly in the areas of Knowledge Graphs (KGs) and Retrieval-Augmented Generation (RAG), offer powerful tools for building context-aware, explainable, and adaptive chatbot systems.

Knowledge Graphs serve as structured representations of relationships between entities, enabling systems to model and reason over complex domains. In the context of satellite mission operations, KGs can map interactions between components, processing algorithms, and historical anomaly data—offering a systemic view of mission behavior. Foundational work such as [1] outlines the theoretical basis and practical applications of KGs in domains requiring interpretability and traceability.

Meanwhile, Retrieval-Augmented Generation represents a paradigm shift in natural language processing, where large pre-trained language models are combined with real-time information retrieval to generate contextually accurate responses. RAG was first introduced in [2] and has since been explored in numerous downstream applications requiring explainability and grounding in external sources.

In high-stakes environments like space operations—where timely and accurate responses to anomalies are vital—combining these technologies holds immense promise. Prior work in this area includes studies on AI-based anomaly detection and resolution, but few efforts have focused on the integration of semantic knowledge representations and generative language models for interactive decision support.

This work presents a demonstrator developed by CS, leveraging Copernicus Sentinel-2 mission data, specifically targeting the image processing chain. The system combines a domain-specific Knowledge Graph with RAG-enhanced chatbot functionality to assist operators in diagnosing anomalies and recommending operational procedures. Through mechanisms such as shadow prompting and targeted retrieval, the system not only identifies potential root causes but also generates actionable, traceable resolution steps.

We present here the application Figure 1 that will be explained in the remaining of the paper.

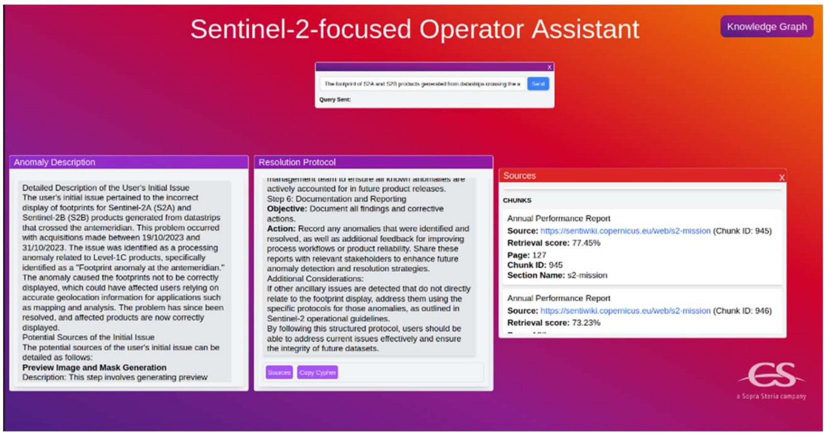

Figure 1 Capture of the application

2. Approach

The central goal of this work is to enable deep, corpus-wide contextual analysis over large collections of documents using large language models (LLMs), beyond the limitations of traditional Retrieval-Augmented Generation (RAG) systems. To achieve this, we construct a domain-specific Knowledge Graph (KG) from the document corpus, which serves both as an intermediate semantic representation and as a reasoning layer to assist the LLM in generating meaningful answers to high-level queries.

2.1. Limitations of Naïve RAG Approaches

Conventional RAG pipelines operate by segmenting documents into smaller chunks and indexing them using vector-based retrieval (e.g., FAISS). At inference time, the LLM receives a limited number (top-k) of these chunks as contextual input. While effective for specific, localised queries, this approach fails to support global analytical tasks that require insight from the entire corpus. Questions like *"What are the main themes?"* or *"What are the conflicting perspectives across documents?"* cannot be fully addressed, as the retrieval mechanism is bound by scoring thresholds and context window limits, often omitting critical information.

2.2. Knowledge Graph Construction

To address this, we introduce a Knowledge Graph-based architecture that abstracts and condenses the document corpus into a graph structure comprising:

- **Entities**: Key concepts, organizations, events, technologies, etc.
- **Relations**: Semantic connections between entities (e.g., “uses,” “produces,” “fails due to”).
- **Document provenance**: Linking entities and edges back to source documents for traceability.

2.3. LLM Integration and Query Processing

The constructed KG is leveraged in two primary ways:

- **Graph-Guided Prompting**: Instead of directly prompting the LLM with raw documents or chunked retrievals, we guide it using summaries derived from the KG. This includes paths, subgraphs, or clusters relevant to the user’s query.
- **Graph-Contextual Reflection**: For analytical queries (e.g., trend detection or theme extraction), the LLM is prompted with an abstracted view of the KG, allowing it to reflect over the whole corpus via a compressed, yet semantically rich, representation

2.4. Data leveraged

In order to enable the reproducibility of our results and approach, we have leveraged solely publicly available data. This decision ensures transparency and accessibility, allowing other researchers and practitioners to validate and build upon our work. The data sources encompass detailed descriptions of the system architecture and accounts of documented anomalies. By focusing on open-access information, we promote a collaborative environment where findings can be independently verified, thereby enhancing the reliability and impact of our research. The use of publicly available data also aids in fostering a broader understanding of the system's behavior and potential points of failure, contributing valuable insights into future improvements and innovations.

- ESA ontology: https://data.esa.int/esado/en/
- SentiWiki: https://sentiwiki.copernicus.eu/web/
- Sentinel 2 anomalies: https://s2anomalies.acri.fr/

In total around 2.5 Go of data was harvested and leveraged to build the KG and RAG’s database, including PDF, webpages, diagrams and anomalies.

2. Methodology

The types of nodes and relationships within the Knowledge Graph, Figure 2, are conceived through a collaborative process, intertwining the developer's deep understanding of the domain with the suggestions generated by the LLM based on sample data. This synergy ensures that the graph structure is both logically sound and semantically enriched, laying a robust foundation for further analytical exploration and enhanced data interoperability.

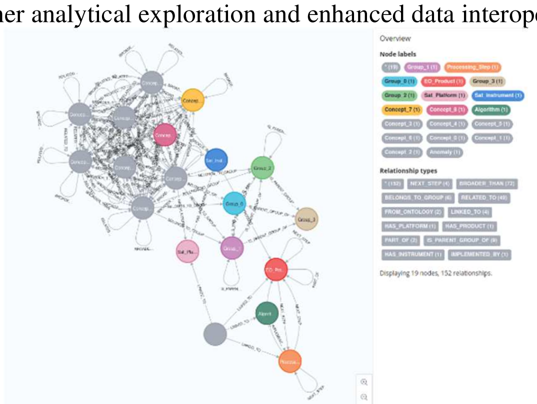

Figure 2 Schema of the knowledge graph

The pipeline for Knowledge Graph construction is visually summarized in Figure 3. The methodology is centred around leveraging large language models (LLMs) to efficiently extract high-quality semantic triplets from a multilingual corpus. The approach consists of two primary phases: document summarization and iterative triplet extraction, followed by a post-processing step.

The process begins with raw textual input, which is passed to a language model to generate a concise summary. This step reduces noise by eliminating non-informative content and lowers computational cost by minimizing token count. Additionally, it ensures language consistency across the corpus, as all summaries are generated in English regardless of the original document language.
The summarized document enters a loop where semantic triplets—structured as (subject, predicate, object)—are extracted one at a time. In each iteration, the LLM is prompted to generate a single novel triplet that is not yet part of the extracted list. This iterative design improves extraction quality by enabling refined prompting and reducing cognitive bias or repetition from prior outputs it also improves the reputability of the knowledge three generated by the LLM. After each iteration, the new triplet is appended to the list and the loop continues.
A conditional decision point checks whether additional triplets remain to be extracted. The process halts either when the model explicitly indicates that no further information can be retrieved or when a predefined maximum number of triplets (empirically set to 100) is reached. Once the list is finalized, it undergoes a clean-up step that includes deduplication, normalization of predicates and entity names, and elimination of semantically redundant or low-value relationships. The cleaned triplets are then used to construct and visualize the Knowledge Graph.

This methodology ensures that the final graph is both compact and semantically rich, facilitating downstream tasks such as document analysis, topic discovery, and question answering.

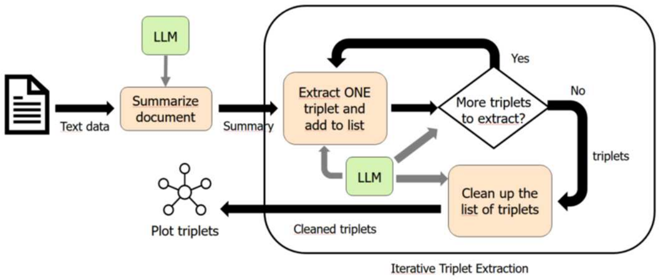

Figure 3 KG generation pipeline

See below Figure 4 the generated subgraph about a given EO product.

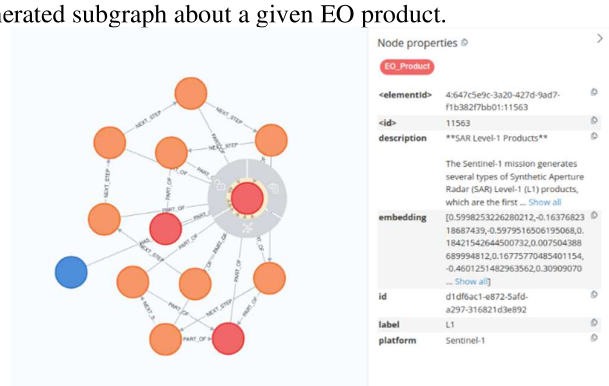

Figure 4 Example of cluster of nodes around an EO product

Based on a new anomaly described as prompt input to the system, in parallel, both the vector database and the content of the Knowledge Graph are utilized to deliver the final answer, Figure 5. The vector database, with its efficient similarity search capabilities, allows for the rapid retrieval of relevant information, while the Knowledge Graph provides a semantically enriched context for these findings. This dual approach ensures that the outputs are both accurate and contextually appropriate. The large language model (LLM) acts as a mediator, integrating insights from both sources to generate comprehensive and precise responses, effectively leveraging the strengths of each to enhance the quality and reliability of the final answer. We empirically tested several configurations and balanced between the data retrieved from the vector data base and the knowledge graph by broadcasting inputs to different configurations of the system and comparing the results.

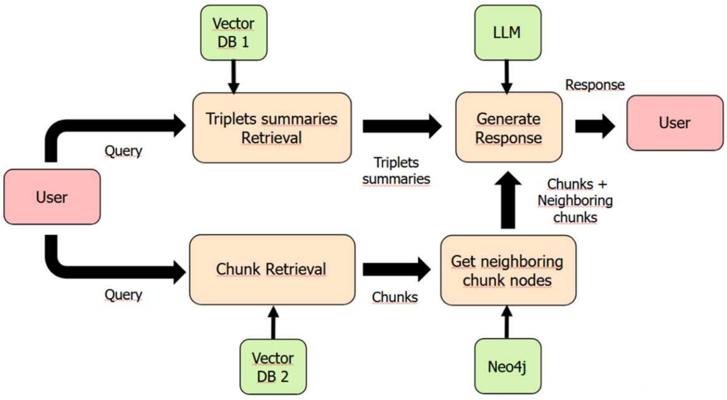

Figure 5 Analysis and Recommendation generation workflow

3. Results

The system developed can be triggered by providing as input anomaly such as: "The footprint of S2A and S2B products generated from datastrips crossing the antemeridian and acquired between 19/10/2023 and 31/10/2023 is not correctly displayed." The following analysis and resolution suggestion are generated in few seconds as illustrated in Figure 6.

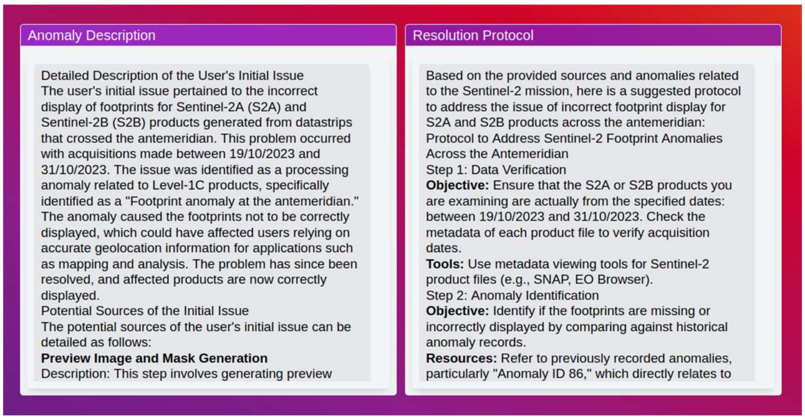

Figure 6 Example of Analysis and Resolution generated for the given anomaly

Here are a sample of the data retrieved from the vector database and the nodes/relations from the graph that have been used, Figure 8 and Figure 7.

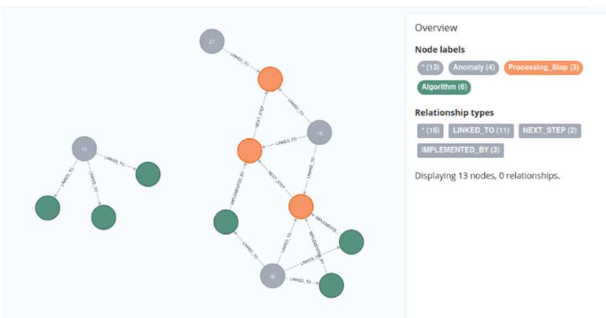

Figure 8 Nodes leveraged for the generation

Figure 7 Document chunks leveraged for the generation

For verification we requested the system to provide 10 answers for each of the 91 anomalies ingested into the KG and the vector database. We performed the same request for each of the 5 anomalies that had not been ingested. The BERT index was quite satisfactory and not knowing the anomaly before do not seems to impact the quality of the answer significantly.

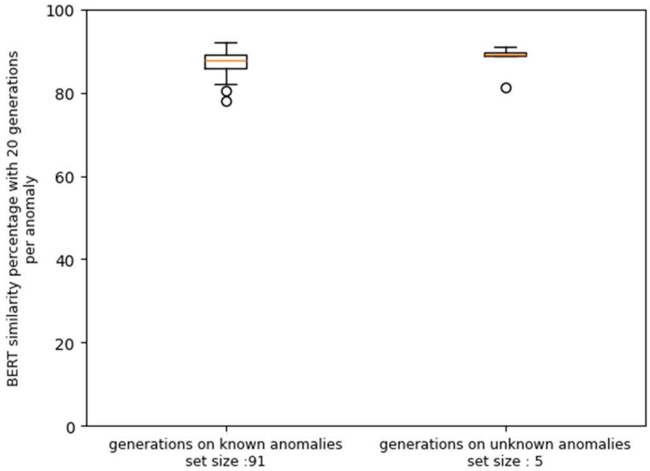

Figure 9 BERT similarity percentage for generations on known and unknown anomalies

5. Discussion

The verification and validation need to obtain significant feedback from operators as while the analysis and recommendation generated seems quite relevant, proper validation by operators is necessary.

The added value of this type of system is seen when limited or no anomaly has been recorded (in our case less than 100) but this complexify the benchmarking of the solution, with the lack of effective details on the resolution protocol followed by operators.

6. Conclusions

This system demonstrated that the combination of Retrieval-Augmented Generation (RAG) and Knowledge Graphs (KGs) can create powerful operator support tools. These tools facilitate and accelerate the work of operators when encountering anomalies, providing them with precise and actionable insights in real-time. The ability to generalize such tools, combined with a deep understanding of the system encapsulated in the KG, ensures that operators can efficiently manage and resolve issues. This type of tool can be developed with limited or no backlog of anomalies. This approach not only enhances operational efficiency but also improves the reliability and accuracy of anomaly resolution, making it an asset in various applications.

References

[1] Hogan et al. (2021), "Knowledge Graphs".
[2] Lewis et al. (2020), "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks".