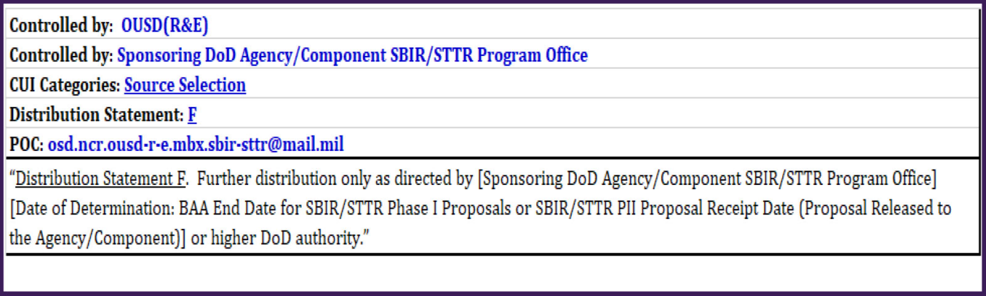
CUI

ATTENTION



Controlled by: OUSD(R&E)

Controlled by: Sponsoring DoD Agency/Component SBIR/STTR Program Office

CUI Categories: Source Selection

Distribution Statement: F

POC: osd.ncr.ousd-r-e.mbx.sbir-sttr@mail.mil

"Distribution Statement F. Further distribution only as directed by [Sponsoring DoD Agency/Component SBIR/STTR Program Office] [Date of Determination: BAA End Date for SBIR/STTR Phase I Proposals or SBIR/STTR PII Proposal Receipt Date (Proposal Released to the Agency/Component)] or higher DoD authority."

ATTENTION

All individuals handling this information are required to protect it from unauthorized disclosure.

Handling, storage, reproduction, and disposition of the attached document(s) must be in accordance with 32 CFR Part 2002 and applicable agency policy.

Access to and dissemination of Controlled Unclassified Information shall be allowed as necessary and permissible to any individual(s), organization(s), or grouping(s) of users, provided such access or dissemination is consistent with or in furtherance of a Lawful Government Purpose and in a manner consistent with applicable law, regulations, and Government-wide policies.

Standard Form 901 (11-18)
Prescribed by GSA/ISOO | 32 CFR 2002

CUI

Small Business Innovation Research(SBIR) Program - Proposal Cover Sheet

Disclaimer

Knowingly and willfully making any false, fictitious, or fraudulent statements or representations may be a felony under the Federal Criminal False Statement Act (18 USC Sec 1001), punishable by a fine of up to $10,000, up to five years in prison, or both.

SBIR Phase I Proposal

Proposal Number:	**F244-0001-0002**
Proposal Title:	**Interactive Knowledge Graphs for Situational Awareness**

Agency Information

Agency Name:	**USAF**
Command:	**AFMC**
Topic Number:	**AF244-0001**

Firm Information

Firm Name:	**Datelite LLC**
Address:	**24310 Wrens Landing CT, Aldie, VA 20105-5939**
Website:	**https://www.datalite.ai**
UEI:	**FN9AE3MN2RB7**
CAGE:	**92C54**
SBA SBC Identification Number:	**002277849**

Firm Certificate

OFFEROR CERTIFIES THAT:

1. It has no more than 500 employees, including the employees of its affiliates.	**YES**
2. Number of employees including all affiliates (average for preceding 12 months)	**12**
3. The business concern meets the ownership and control requirements set forth in 13 C.F.R. Section 121.702.	**YES**
4. Verify that your firm has registered in the SBAS Company Registry at www.sbir.gov by providing the SBC Control ID# and uploading the registration confirmation PDF: **Supporting Documentation:** • SBC_002277849.pdf	**SBC_002277849**
5. It has more than 50% owned by a single Venture Capital Owned Company (VCOC), hedge fund, or	**NO**

private equity firm

6. It has more than 50% owned by multiple business concerns that are VOCs, hedge funds, or private equity firms? **NO**

7. The birth certificates, naturalization papers, or passports show that any individuals it relies upon to meet the eligibility requirements are U.S. citizens or permanent resident aliens in the United States. **YES**

8. Is 50% or more of your firm owned or managed by a corporate entity? **NO**

9. Is your firm affiliated as set forth in 13 CFR Section 121.103? **NO**

10. It has met the performance benchmarks as listed by the SBA on their website as eligible to participate **N/A**

11. Firms PI, CO, or owner, a faculty member or student of an institution of higher education **NO**

12. The offeror qualifies as a:

[**X**] Socially and economically disadvantaged SBC

[**X**] Women-owned SBC

[ ] HUBZone-owned SBC

[ ] Veteran-owned SBC

[ ] Service Disabled Veteran-owned SBC

[ ] None Listed

13. Race of the offeror:

[ ] American Indian or Alaska Native

[ ] Native Hawaiian or Other Pacific Islander

[**X**] Asian

[ ] White

[ ] Black or African American

[ ] Do not wish to Provide

14. Ethnicity of the offeror: **NON-HISPANIC**

15. It is a corporation that has some unpaid Federal tax liability that has been assessed, for which all judicial and administrative remedies have not been exhausted or have not lapsed, and that is not being paid in a timely manner pursuant to an agreement with the authority responsible for collecting the tax liability: **FALSE**

16. Firm been convicted of a fraud-related crime involving SBIR and/or STTR funds or found civilly liable for a fraud-related violation involving federal funds: **NO**

17. Firms Principal Investigator (PI) or Corporate Official (CO), or owner been convicted of a fraud-related crime involving SBIR and/or STTR funds or found civilly liable for a fraud-related violation involving federal funds: **NO**

Signature:

Printed Name	Signature	Title	Business Name	Date
Anand Thiagarajan	ANAND THIAGA RAJAN	EVP	Anand	08/19/2024

Audit Information

Summary:

Has your Firm ever had a DCAA review?**NO**

VOL I - Proposal Summary

Summary:

Proposed Base Duration (in months):	**6**

Technical Abstract:

Datalite is proposing to develop a state-of-the-art solution to enhance the Air Force's situational awareness through interactive and adaptive Knowledge Graphs (KGs). Leveraging expertise in AI, Knowledge Graphs, and Information Retrieval (IR), Datalite's approach addresses two core elements of KG modification: identifying areas within the graph to change and performing the modifications. Traditional methods rely on user inputs with minimal automation, limiting efficiency in dynamic, time-sensitive environments. Datalite's solution proposes a more advanced, AI-driven framework for effective and efficient KG management.

The proposed methodology focuses on applying both generative and non-generative AI/ML techniques. Generative models, such as Large Language Models (LLMs), are intended to automate tasks including classifying manual edits into thematic "Edit Classes" and generating queries to locate graph elements needing similar updates. Non-generative models utilize embeddings (e.g., node2vec) to identify nodes requiring edits by finding semantically similar nodes to user-selected sites, enabling rapid, index-based retrieval followed by detailed evaluation.

The research in Phase I targets two key areas: Edit Class Formation and Query Generation. Edit Class Formation involves using LLMs to generate descriptive narratives for groups of similar user-initiated edits, allowing for automation of repetitive tasks. Query Generation, combining vector embeddings and generative AI, facilitates the rapid identification of graph locations requiring edits based on established Edit Classes. These techniques promise substantial time savings and improved accuracy in KG modifications, directly supporting Air Force objectives in Trusted AI and Autonomy.

The solution employs automated metrics (e.g., accuracy, F1 score) and manual human review for evaluation. Phase I research is intentionally oriented toward LLM-based techniques due to limited access to classified data, with future phases incorporating fine-tuned ML approaches as additional data becomes available. Ultimately, Datalite's innovative framework aims to streamline KG modification workflows, offering the Air Force a robust tool for enhanced situational awareness and decision-making in critical operations.

Anticipated Benefits/Potential Commercial Applications of the Research or Development:

Datalite’s research focuses on meeting the situational awareness and data analysis needs of the Department of Defense (DoD), other Federal agencies, and private sector markets. This technology, grounded in adaptive and dynamic knowledge graphs, is designed to enhance real-time data interactions, allowing users to interact with and update data to improve situational awareness, detect threats, and conduct patterns of life analysis.

In the DoD, situational awareness and real-time data analysis are critical for effective operations, especially in high-stakes, time-constrained environments like combat or intelligence. Traditional knowledge graph systems lack adaptability and rely heavily on manual updates, which slows down decision-making. Datalite’s solution allows real-time interaction and updates to the knowledge graph, incorporating user feedback to automatically suggest additional relevant connections and corrections. This capability aligns with the DoD’s priorities in Trusted AI and Autonomy, supporting applications such as targeting and intelligence operations.

For other Federal agencies, such as Homeland Security and law enforcement, the technology offers similar benefits by aiding in threat detection, data fusion, and information retrieval. These applications align well with the growing demand for AI-driven solutions that can provide insights and improve response times.

In the private sector, financial markets and cybersecurity industries face challenges with large, complex datasets where adaptive knowledge graphs can support anomaly detection, fraud analysis, and strategic data management. The financial services market, in particular, is actively seeking innovative solutions to analyze interconnected data for real-time insights.

This research will help users make manual edits to both improve data quality in a knowledge graph and adapt to changes in a dynamic knowledge graph environment, supporting faster, more effective situational awareness, more accurate pattern of life analysis, precise and correct threat detection, and optimized targeting operations in time-constrained environments.

Attention:

Disclaimer: For any purpose other than to evaluate the proposal, this data except proposal cover sheets shall not be disclosed outside the Government and shall not be duplicated, used or disclosed in whole or in part, provided that if a contract is awarded to this proposer as a result of or in connection with the submission of this data, the Government shall have the right to duplicate, use or disclose the data to the extent provided in the funding agreement. This restriction does not limit the Government's right to use information contained in the data if it is obtained from another source without restriction. This restriction does not apply to routine handling of proposals for administrative purposes by Government support contractors. The data subject to this restriction is contained on the pages of the proposal listed on the line below.

Addition:

Enter the page numbers separated by a space of the pages in the proposal that are considered proprietary:

0

List a maximum of 8 Key Words or phrases, separated by commas, that describe the Project:

GCN, Node2Vec, Vector Embedding, Knowledge Graph, Edit Class

VOL I - Proposal Certification

Summary:

1. At a minimum, two thirds of the work in Phase I will be carried out by your small business as defined by 13 C.F.R Section 701-705. The numbers for this certification are derived from the budget template. To update these numbers, review and revise your budget data. If the minimum percentage of work numbers are not met, then a letter of explanation or written approval from the funding officer is required. Please note that some components will not accept any deviation from the Percentage of Work (POW) minimum requirements. Please check your component instructions regarding the POW requirements.	**YES**
Firm POW	**100%**
Subcontractor POW	**0%**
2. Is primary employment of the principal investigator with your firm as defined by 13 C.F.R Section 701-705?	**YES**
3. During the performance of the contract, the research/research and development will be performed in the United States.	**YES**
4. During the performance of the contract, the research/research and development will be performed at the offerors facilities by the offerors employees except as otherwise indicated in the technical proposal.	**YES**
5. Do you plan to use Federal facilities, laboratories, or equipment?	**NO**
6. The offeror understands and shall comply with export control regulations.	**YES**
7. There will be ITAR/EAR data in this work and/or deliverables.	**NO**
8. Has a proposal for essentially equivalent work been submitted to other US government agencies or DoD components?	**NO**
9. Has a contract been awarded for any of the proposals listed above?	**NO**
10. Firm will notify the Federal agency immediately if all or a portion of the work authorized and funded under this proposal is subsequently funded by another Federal agency.	**YES**
11. Are you submitting assertions in accordance with DFARS 252.227-7017 Identification and assertions use, release, or disclosure restriction?	**NO**
12. Are you proposing research that utilizes human/animal subjects or a recombinant DNA as described in DoDI 3216.01, 32 C.F.R. Section 219, and National Institutes of Health Guidelines for Research Involving Recombinant DNA of the solicitation:	**NO**
13. In accordance with Federal Acquisition Regulation 4.2105, at the time of proposal submission, the required certification template, "Contractor Certification Regarding Provision of Prohibited Video Surveillance and Telecommunications Services and Equipment" will be completed, signed by an authorized company official, and included in Volume V: Supporting Documents of this proposal.	**YES**

NOTE: Failure to complete and submit the required certifications as a part of the proposal submission process may be cause for rejection of the proposal submission without evaluation.

14. Are teaming partners or subcontractors proposed?	**NO**
15. Are you proposing to use foreign nationals as defined in 22 CFR 120.16 for work under the proposed effort?	**NO**
16. What percentage of the principal investigators total time will be on the project?	**80%**
17. Is the principal investigator socially/economically disadvantaged?	**YES**
18. Does your firm allow for the release of its contact information to Economic Development Organizations?	**YES**

VOL I - Contact Information

Principal Investigator

Name:	**Mr. Anand Thiagarajan**
Phone:	**(571) 499-0845**
Email:	**anand@datalite.ai**
Address:	**24310 Wrens Landing CT, Aldie, VA 20105 - 5939**

Corporate Official

Name:	**Pallavi Anand**
Phone:	**(240) 477-3991**
Email:	**Pallavi.Goel@datalite.ai**
Address:	**24310 Wrens Landing CT, Aldie, VA 20105 - 5939**

Authorized Contract Negotiator

Name:	**Mr. Anand Thiagarajan**
Phone:	**(571) 499-0845**
Email:	**anand@datalite.ai**
Address:	**24310 Wrens Landing CT, Aldie, VA 20105 - 5939**

Form Generated on 11/06/2024 12:51:20 AM

Volume 2: Technical Volume
Interactive Knowledge Graphs for Situational Awareness

Glossary of terms

- **Knowledge Graph (KG)**: A data structure that represents information as a network of entities (nodes) and their relationships (edges), often used for complex data integration, retrieval, and analysis.
- **Human-in-the-Loop:** A process design that combines automated and human contributions, allowing human oversight or intervention to improve the accuracy and reliability of automated systems.
- **Edit Class:** A category of similar edits in a knowledge graph, inferred from repeated manual changes, which can be used to automate future edits within that class.
- **Graph Convolutional Network (GCN):** A type of neural network designed to operate on graph-structured data, commonly used for tasks like node classification and link prediction in static graphs.
- **Node2Vec:** An unsupervised learning algorithm that generates node embeddings by performing random walks on a graph to capture relationships between nodes.
- **Vector Embeddings:** Numerical representations of entities (e.g., graph nodes) in a high-dimensional space, allowing for similarity comparisons. Used here to identify similar nodes for automated edits.
- **F1 Score:** A performance metric that combines precision (correctness of positive predictions) and recall (coverage of actual positives) to assess model accuracy, particularly useful in evaluating classification or retrieval systems.
- **Cypher/SPARQL Query:** A structured query language used for retrieving and modifying data in knowledge graphs (Cypher for Neo4j, SPARQL for RDF-based KGs).
- **Retrieval-Augmented Generation (RAG):** A machine learning framework where retrieval of relevant information from a database augments an LLM's response, enhancing its output with specific data.
- **Generative AI:** AI techniques, particularly using large language models (LLMs), that generate new content or data, such as text, based on patterns learned from training data.

1. Identification and Significance of the Problem or Opportunity

Air Force (AF) operations face many challenges in enabling real-time user interaction with dynamic knowledge graphs to improve situational awareness, pattern analysis, threat detection, and targeting in time-constrained environments.

Summary of key issues

- Current AI/ML methods are used to structure and store data within knowledge graphs, but these methods are not fully trusted by human analysts. Errors, incomplete structuring, and unintended correlations introduced by AI models often require analysts to manually correct or supplement the data.
- Presently, analysts must manually interact with knowledge graphs to correct or update data, which is time-consuming and becomes infeasible in critical, time-constrained environments.
- Existing AI/ML techniques to predict edits for a graph may be too slow to work with dynamic graphs, and where rapid results are needed.
- Many data pipelines are fully automated for speed but lack human-in-the-loop capabilities that will also leverage domain knowledge and intelligence of human experts.
- In addition to manual edits being slow and possibly repetitive, manual edits may help find conflicts and gaps in the knowledge graph that a user would not otherwise be aware of, so would not manually correct without automated assistance in highlighting them.

Datalite believes that there is an opportunity to add human-in-the-loop in a new and productive way, by generating edits based on a small set of human edits, particularly if edits tend to share structure in which case we can consider classes of edits, and the classes may be inferred or generated.

Normally, established techniques for predicting graph changes, particularly well-established link prediction approaches are done using static methods such as training a GCN on the graph. Innovative solutions are needed in part due to the dynamic nature of the graphs, which will not allow retraining a GCN on a moderate-to-large graph as “daily-weekly” updates are made per the RFP Q&A section.

Overall, automated techniques that leverage the intelligence of human users by examining the edits they choose to make can bring human guidance into a KG system, correct data, deconflict information, close gaps, and accelerate KG curation to a point it becomes useful in dynamic environments such as situational awareness, pattern of life analysis, threat detection and targeting operations.

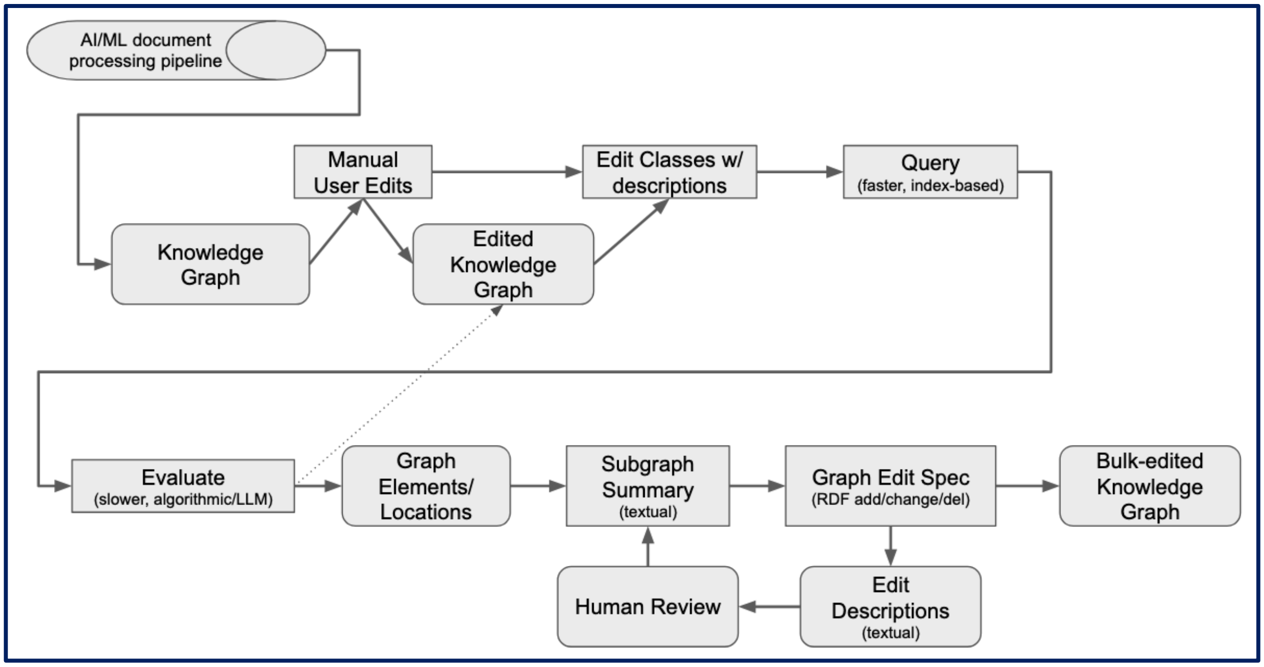
2. Phase I Technical Objectives

Datalite brings an innovative and in-depth understanding of the Air Force's challenges, as well as expertise in AI and Knowledge Graphs (KG) which guides our research approach and helps ensure that our research will advance the state of the art.

Information Retrieval vs Graph Editing

Two elements are needed to alter a KG. First, the area or elements of the graph to change must be identified. Then the change at that location must be done. This first sub-problem is in the area of Information Retrieval (IR).

Our team includes database experts who have worked on multiple graph and non-graph database products, and we can further assert that IR is a two-phase process: first perform efficient (typically index-based) retrieval of candidates, then perform less efficient "filtering" or evaluation of candidates using slower techniques. The flow for graph edits is illustrated in the exhibit below:


AI/ML document processing pipeline
Manual User Edits
Edit Classes w/ descriptions
Query (faster, index-based)
Knowledge Graph
Edited Knowledge Graph
Evaluate (slower, algorithmic/LLM)
Graph Elements/ Locations
Subgraph Summary (textual)
Graph Edit Spec (RDF add/change/del)
Bulk-edited Knowledge Graph
Human Review
Edit Descriptions (textual)

Some processes, such as querying a KG with a Cypher or SPARQL query are existing technology that require no innovation or research. Other areas, such as classifying repeated human-initiated edits into "Edit Classes" that are repeatable, require innovation and research.

Generative AI vs non-generative AI/ML

Because recent AI work is focused on LLMs and other generative technologies, but the problem space here is that of symbolic AI (that is, knowledge graphs and other curated, explicit symbolic information), it is important to consider what parts of the problem flow are generative vs not.

We propose to research three generative elements in the above graph editing process:

1. Generating a (textual) description of a "class" of manual, user-initiated graph changes
2. Generating a query to find graph elements or areas that should change
3. Generating a set of graph edits at each identified location, to implement the changes based on prior user edits

A non-generative element in the above flow may include Query generation. A generative approach to Query generation using LLMs is possible, and we propose to evaluate that. However, a faster and possibly more effective approach is to use vector embeddings for every node in the graph to find semantically similar graph locations to make edits, based on human-selected edit locations for the edits in the class. This would be the index-based phase, still followed by a more computationally intensive evaluation phase that may use LLMs.

Scope of Research

We propose to study two of the above processing areas: Edit Class Formation and Query Generation. These two areas will work together, as the representation of the Edit Class can be used to generate an appropriate query.

In addition, these areas can separately or together vastly improve the efficiency of a human who is manually correcting issues in a knowledge graph via similar, repetitive edits; that is, other areas need not be researched in the Phase I effort to produce immediate value to the Air Force.

Here are these research areas, mapped to the flow diagram above, and our suggested approach for each.

Research Topic	Proposed Work	Evaluation Criteria
Edit Class formation	Select prompts and LLMs that will take prior edits as input and generate descriptions of the overall Edit Class, meaning a thematically grouped description of edits that would have to be done manually, but are similar and can be automated. This will produce narrative text describing the Edit Class and which is suitable for both further LLM processing and human review.	Accuracy and F1 score. Manual human evaluation of each Edit Class description, to determine if the narrative text accurately and completely describes the nature of a related group of human-made edits. Matrix of Prompt x LLM selection with accuracy measures for each

		Prompt/LLM combination tested.
Query generation (non-generative)	Apply two non-trained graph encoding algorithms that each create a vector for every node in the graph: node2vec and LLM-based narrative encoding using domain knowledge (more on this below). Use a vector database (e.g. Postgres+pgVector or Qdrant) to retrieve semantically similar nodes to the centroid of human-user edit graph sites; these similar nodes will be candidate sites for future, automated edits.	F1-score of the accuracy of the selection of future edit sites.
Query generation (generative)	Evaluate LLM techniques to create a Cypher or SPARQL query based on the narrative textual description of each Edit Class.	F1-score of the accuracy of the selection of future edit sites.

Note that this Phase I work is biased somewhat toward LLM based techniques due to the lack of availability of classified data. Techniques that train an ML system, e.g. using GraphSAGE or GCNs to create graph embeddings, and which work on non-representative data may not hold up on the actual data.

Therefore, we propose to use LLMs which are pre-trained with all needed human knowledge. In a later phase, trained ML approaches to graph embedding, and fine-tuned LLMs that "know" about the Air Force data and domain may be evaluated or added with relatively little additional work, and leveraging the testing harnesses and techniques developed in this SBIR Phase I effort.

3. Identification Phase I Statement of Work

Datalite shall work on three research topics. All the work will be done in the Datalite office using its staff. Each research topic may take two months of duration. Each of the 3 research topics is discussed in detail.

1. Edit Class Formation

The work on Edit Class Formation will be to create a process that reads a set of edits and produces a definition of one or more Edit Classes, where an Edit Class represents the theme or similar task among a group of edits thus capturing a user's underlying intent based on their manual edits so far. Equivalently, we can think of an Edit Class as a representation of the type of data issue being corrected. For simplicity, we will assume all Edits in a set are in the same class, which eventual users of a system can ensure by making edits specific to one class and then requesting automated help with the rest of that Class of edit.

Example:

If a NLP-based data pipeline incorrectly interprets "Emir" or "Naqeeb" as given names rather than paramilitary titles (both implying "commander") then a repeated edit may be to find Person entities with given name of Emir or Naqeeb, remove the given name property or set to UNKNOWN, add a presumed Role link to a new Role entity, and set the Role.type to "CMDR."

The Edit Class description will be much like the description above in this proposal:

"For all Person entities (pers) with given name equal to "Emir" or "Naqeeb," remove the pers.givenName property. Add a presumedRole link from pers to a new Role entity (new_role), and set the Role.type of new_role to "CMDR."

The research will include:

- Determining a data representation for an edit and writing up a set of manual human edits in this data representation. An edit is likely to be represented by a main node plus a set of RDF adds and deletes, plus contextual information about the data around the main node, which provides context. This test data will drive evaluation.
- Generating a set of four possible Prompts that instruct the LLM how to interpret a set of manual edits and create an Edit Class. Few-shot prompting, where examples are included in the prompt, will be utilized.
- Selecting three LLMs to test. As of this writing, these may be OpenAI 4o, Anthropic Claude 3.5, and Llama 3.1 8b as they are high-end models, but avoid the expensive GPT o1 and resource intensive Llama 3.1 405B options. We will re-evaluate a set of capable but fast and affordable options at the start of the effort, as these models evolve quickly.
- A test harness will be developed that uses every combination of LLM + Prompt and predicts edits in the knowledge graph based on prior human edits.
- Measuring accuracy and F1 score for each of the 12 Prompt + LLM combinations.

2. Query generation (vector based)

Vector databases provide semantic similarity queries using vectors. In this approach, every item (nodes in the graph in our case) is assigned a vector that semantically represents the node. Once vectors are populated, the vector database will tell us which nodes are similar to any given node (or vector centroid of a set of nodes). This allows us to take existing human-made edit locations (nodes) and determine what other nodes are similar, and which, therefore may be likely edit locations for future automated edits.

The work on vector-based Query Generation will be in choosing an algorithm to tag every node in the graph with a vector and evaluating the accuracy of vector-based retrieval of candidate edit locations in the knowledge graph. The Query will be a basic vector-similarity query using the human-edited nodes as input (likely their centroid).

Creating vectors for nodes in a graph is known as the "graph embedding problem," for which there are many known best practices and approaches. As a practical matter, we will focus on approaches that do not require training an ML solution using real data, as we will do this work in an unclassified environment that does not have access to real data. Training an ML solution on synthetic data, or data in another domain will not be likely to be predictive, so is a less useful approach.

Restricting non-trained approaches, we will use node2vec and LLM-based vector generation to associate a vector with every node in our test knowledge graph.

node2vec

node2vec is a well-understood algorithm that uses random walks to take a node in a graph and turn it into a sequence of nearby nodes, which are analogous to words in a sentence. (The technique is inspired by the successful text word2vec approach).

LLM vector embeddings

The LLM-based approach will be to use domain knowledge to describe a node of a particular type in the graph using English language. To do this, three code snippets will be written, one for each of three node types in the knowledge graph, and these code snippets will produce an English summary of the node.

All LLMs use a latent semantic space to encode information when answering questions or generating content, and all provide access to the embedding vector that represents input text in this latent semantic space. In particular, the popular LangChain library has a model-independent API to generate embeddings from text, which is: embeddings_model.embed_text(text). We will use this call to convert the node textual description for each node to a vector.

Example

For a graph that has Facilities, Persons and Roles, and a facility is the CVS on the corner of 1st St and Main St. in Anytown, USA, the textual description of the Facility may be: *"CVS_192 is a Facility. CVS_192 is called "CVS". CVS_192 has the address "188 Main St, Anytown USA". Bob_Jones is a Person. Bob_jones name is "Bob Jones". Bob_Jones works at CVS_192."*

It may be evident that underlying RDF triples in the knowledge graph, which are already in subject-predicate-object form, can be trivially translated into English text. Using English text enables LLMs, which are trained on English text, to process the information with better accuracy.

Note that in the future GCNs and other ML-trained graph encoding approaches can be tested using the same test harness and approach as we use for node2vec and LLM-based encoding.
GCNs are a classic and powerful way to do this, but too slow to meet Air Force needs. We suggest future work may use GraphSAGE for node embeddings, as it uses a local area around a node to create an embedding, so changes to a graph only necessitate re-computing embeddings in a

subgraph within a set number of hops from the changed nodes. Evaluating GCNs or GraphSAGE is not proposed in this scope of work, but the work proposed here sets up future evaluations of that kind.

3. Query generation (generative, text based)

We will also evaluate using LLMs more directly for query generation, rather than using LLMs for vector generation, and then querying using vector similarity.

Unlike the vector-based approach, no embeddings are needed. Instead, an LLM will be used to convert the English text description of the Edit Class into a structured Cypher query. (Recall that textual descriptions of human edits will be produced under the first research topic area: Edit Class Formation as described above.) Many knowledge graph systems, including neo4j, will execute a Cypher query. In the future, the same technique can be used to produce SPARQL or other query specifications as well, but that is not within the proposed scope here.

This approach is promising because it is central to RAG-based chat systems, and is therefore mature and well-understood, and improving rapidly. RAG is “retrieval augmented generation” where real-time or proprietary data is added to an LLM chat request to “augment” the chat request with specific context. Often, the proprietary data must come from a database, so LLMs are used to create database queries that retrieve required information. Neo4j provides “cypherchain.run()” as a function that turns text prompts into Cypher queries.

Evaluation of query approaches

For all three approaches – the two vector-based query approaches using node2vec and LLM embeddings and the LLM-based Cypher query generation – we will compute the F1 score by comparing the returned candidate nodes where the system is saying an edit may need to take place with our test data that lists the actual locations where that Edit Class should be applied.

We will compute accuracy, precision and F1 scores for each approach, which will help us understand the relative merits of all three approaches.

4. Related Work

None.

5. Relationship with Future Research or Research and Development

Datalite proposes this research as it is innovative and designed to produce immediate benefits and also be a foundation for future investigations.

The overall likely data flow that combines Classification, Information Retrieval, Query Generation, and Edit Generation is based on our team’s expertise in information retrieval, databases, knowledge graphs, general AI, and LLMs. This overall data flow then suggests particular research activities that are achievable within the SBIR Phase I budget and period of performance limitations.

The research includes a structured, disciplined approach to testing the selected components of what we believe to be an intuitive and effective overall approach. Each tested area can also be used immediately and in isolation in two ways name Specific Utility and Generic Utility.

Specific Utility

Edit Classification alone is useful because it produces a human-readable summary that can be reviewed by the human making the edits. If the Edit Class narrative description is incorrect, the human can correct it using English text edits, without requiring specialized ML skills. Further, the same technology that generates the Edit Class can evaluate if a human-edited description is consistent with the edits or not, as a sanity check.

Vector embeddings of the nodes feature is useful because vector-based semantic similarity search of nodes can be used in isolation, or in combination with other techniques to find interesting or similar nodes in a graph, both to drive bulk editing and for other reasons, such as target selection. If one target is found, other “similar” locations may be considered as targets, etc.

Query generation based on Edit Classes is useful, even in isolation, if a user hand-edits the nature of the bulk edit they wish to make globally. While manual edits may still be slower than desired, just using queries to evaluate the edit locations can vastly speed up the time to result in a large set of edits by directing the user to the right places where he or she can manually edit the data. Further, the text-oriented query of a graph is useful in its own right for tasks other than editing. Again, considering target selection, the nature of the targets desired can be entered as text, and this can become a knowledge graph query showing matching nodes in the graph.

General Utility

This research also addresses broad themes of interest to the Air Force and other potential users. The integration of LLM-based, generative technology to work with more rigorous, curated knowledge graph technology is useful in many domains. Knowledge graphs are “symbolic AI” and have higher-quality, human-curated information, and seamlessly integrate with open data sources and many existing technologies, including inference and visualization, LLMs are “statistical AI” and are growing and changing more rapidly this year and beyond. Innovatively combining statistical and symbolic AI may yield results far beyond this particular use case and be of particular use to organizations like the Air Force that have invested in knowledge graph technology already.

This research will help users make manual edits to both improve data quality in a knowledge graph and adapt to changes in a dynamic knowledge graph environment, supporting faster, more effective situational awareness, more accurate pattern of life analysis, precise and correct threat detection, and optimized targeting operations in time-constrained environments.

6. Commercialization Strategy

Datalite's strategy for commercializing its interactive knowledge graph technology focuses on meeting the situational awareness and data analysis needs of the Department of Defense (DoD), other Federal agencies, and private sector markets. This technology, grounded in adaptive and dynamic knowledge graphs, is designed to enhance real-time data interactions, allowing users to interact with and update data to improve situational awareness, detect threats, and conduct patterns of life analysis.

Market Need and Opportunity

In the DoD, situational awareness and real-time data analysis are critical for effective operations, especially in high-stakes, time-constrained environments like combat or intelligence. Traditional knowledge graph systems lack adaptability and rely heavily on manual updates, which slows down decision-making. Datalite's solution allows real-time interaction and updates to the knowledge graph, incorporating user feedback to automatically suggest additional relevant connections and corrections. This capability aligns with the DoD's priorities in Trusted AI and Autonomy, supporting applications such as targeting and intelligence operations.

For other Federal agencies, such as Homeland Security and law enforcement, the technology offers similar benefits by aiding in threat detection, data fusion, and information retrieval. These applications align well with the growing demand for AI-driven solutions that can provide insights and improve response times.

In the private sector, financial markets and cybersecurity industries face challenges with large, complex datasets where adaptive knowledge graphs can support anomaly detection, fraud analysis, and strategic data management. The financial services market, in particular, is actively seeking innovative solutions to analyze interconnected data for real-time insights.

Market Size

The global knowledge graph market is projected to grow significantly, driven by rising demand in both government and commercial sectors. The defense AI market alone is expected to reach billions of dollars within the next decade as AI integration becomes central to defense strategies. Similarly, the financial sector's investment in AI tools is expanding, with data analytics and fraud detection expected to drive substantial market growth.

Commercialization Path

Phase III of the development will involve tailoring this technology for specific applications within these markets, focusing on seamless integration with existing systems in the DoD, Federal agencies, and private enterprises. Datalite plans to pursue partnership opportunities for pilot programs within these sectors, demonstrating the technology's efficacy in real-world scenarios and further refining its application for each unique environment. By leveraging dual-use pathways, Datalite aims to provide a versatile tool adaptable across various high-demand markets.

7. Key Personnel

Name:	Anand Thiagarajan
Role:	Principal Investigator
Education:	• Bharathiar University,Bachelor of Science in Computer Science – 1998 • Bharathiar University,Master of Science in Computer Science – 2001 • Case Western Reserve University, Master of Business Administration -2012
Relevant Experience	**Technical Architect, Datalite since 2017.** • Designed and developed NoSQL database and Graph database solutions for commercial, state, and federal clients. **Technical Head, HTC Global Services 2007-2017** • Designed and developed NoSQL databases and Graph databases for near real time data display and network analysis.

Name:	Damon Feldman
Role:	Research Guide/Reviewer
Education:	• University of Chicago, BS Math/Computer Science - 1989 • Tulane University, Ph.D. Computer Science - 1996
Relevant Experience	**Senior Director, Consulting, MarkLogic Corp, 2008-2022** • Responsibilities included implementing and overseeing XML, JSON and RDF-based graph data systems for MarkLogic customers. **VP Customer Engineering, Dgraph Labs, 2023-2024** • Responsibilities included supporting customers with new and existing solutions using the Dgraph property graph and knowledge graph system, and augmenting graph-based systems with AI technology and LLM use.

8. Foreign Citizens

Not Applicable.

9. Facilities/Equipment

Personal laptops, Cloud Computing and graph databases

10. Prior, Current, or Pending Number of Similar Proposals or Awards

None.

11. Prior, Identification and Assertion of Restrictions on the Government's Use, Release, or Disclosure of Technical Data or Computer Software.

None.

DSIP Defense SBIR/STTR Innovation Portal

Cost Volume

SBIR Phase I Proposal

Proposal Number	F244-0001-0002
Topic Number	AF244-0001
Proposal Title	Interactive Knowledge Graphs for Situational Awareness
Date Submitted	11/06/2024 12:51:17 AM

Firm Information

Firm Name	Datelite LLC
Mail Address	24310 Wrens Landing CT, Aldie, Virginia, 20105
Website Address	https://www.datalite.ai
UEI	FN9AE3MN2RB7
Cage	92C54

Total Dollar Amount for this Proposal	$135,996.00

Base Year	$133,476.00
Year 2	$2,520.00
Technical and Business Assistance(TABA)- Base	$0.00
TABA- Year 2	$0.00

Base Year Summary

Total Direct Labor (TDL)	$132,600.00
Total Direct Material Costs (TDM)	$0.00
Total Direct Supplies Costs (TDS)	$0.00
Total Direct Equipment Costs (TDE)	$876.00
Total Direct Travel Costs (TDT)	$0.00
Total Other Direct Costs (TODC)	$0.00
G&A (rate 1%) x Base ()	$0.00
Total Firm Costs	$133,476.00
Subcontractor Costs	
Total Subcontractor Costs (TSC)	$0.00
Cost Sharing	-$0.00
Profit Rate (0%)	$0.00
Total Estimated Cost	$133,476.00
TABA	$0.00

Year 2 Summary

Total Direct Labor (TDL)	$2,520.00
Total Direct Material Costs (TDM)	$0.00
Total Direct Supplies Costs (TDS)	$0.00

Total Direct Equipment Costs (TDE)	$0.00
Total Direct Travel Costs (TDT)	$0.00
Total Other Direct Costs (TODC)	$0.00
G&A (rate 1%) x Base ()	$0.00
Total Firm Costs	$2,520.00
Subcontractor Costs	
Total Subcontractor Costs (TSC)	$0.00
Cost Sharing	-$0.00
Profit Rate (0%)	$0.00
Total Estimated Cost	$2,520.00
TABA	$0.00

Base Year

Direct Labor Costs

Category / Individual-TR	**Rate/Hour**	**Estimated Hours**	**Fringe Rate (%)**	**Fringe Cost**	**Cost**
Software Developer/ Principal Investigator	$120.00	760	0	$0.00	$91,200.00
Computer and Information Research Scientist/ Research Guide/Reviewer (Damon Feldman)	$230.00	180			$41,400.00

Subtotal Direct Labor (DL)	$132,600.00
Labor Overhead Cost	$0.00
Total Direct Labor (TDL)	**$132,600.00**

Direct Equipment Costs

Neo4J	$876.00
Total Direct Equipment Costs (DE)	**$876.00**

G&A (rate 1%) x Base ()	$0.00
Cost Sharing	-$0.00
Profit Rate (0%)	$0.00
Total Estimated Cost	$133,476.00
TABA	$0.00

Year 2

Direct Labor Costs

Category / Individual-TR	**Rate/Hour**	**Estimated Hours**	**Fringe Rate (%)**	**Fringe Cost**	**Cost**
Computer and Information Research Scientist/ Principal Investigator	$120.00	20			$2,400.00

Subtotal Direct Labor (DL)	$2,400.00
Labor Overhead (rate 5%) x (DL)	$120.00
Total Direct Labor (TDL)	**$2,520.00**

Direct Equipment Costs

Neo4J	$0.00
Total Direct Equipment Costs (DE)	**$0.00**

G&A (rate 1%) x Base ()	$0.00
Cost Sharing	-$0.00
Profit Rate (0%)	$0.00
Total Estimated Cost	$2,520.00
TABA	$0.00

Explaratory Material Relating to the Cost Volume

The Official From the Firm that is responsible for the cost breakdown

Name: Anand Thiagarajan
Phone: (571) 499-0845
Phone: anand@datalite.ai
Title: Proposal Owner

If the Defence Contracting Audit Agency has performed a review of your projects within the past 12 months, please provide: No
Select the Type of Payment Desired: Partial payments

Cost Volume Details

Direct Labor

Base

Category	Description	Education	Yrs Experience	Hours	Rate	Fringe Rate	Total
Software Developer	Principal Investigator	Master's Degree	25	760	$120.00	0	$91,200.00
Computer and Information Research Scientist	Research Guide/Reviewer	PhD	25	180	$230.00		$41,400.00

Are the labor rates detailed below fully loaded? **YES**

Please explain any costs that apply.
The software developer develops the solution based on the proposal. The research guide reviews the solution and guides the developer. At a minimum, these two resources are required.

Provide any additional information and cost support data related to the nature of the direct labor detailed above.
Graph Data Developers are difficult to find in the market. The developer must be experienced in Graph Databases, AI, Machine Learning, and Cypher and SPARQL. The guide should be a Graph data expert.

Direct Labor Cost ($): $132,600.00

Year2

Category	Description	Education	Yrs Experience	Hours	Rate	Fringe Rate	Total
Computer and Information Research Scientist	Principal Investigator	Master's Degree	25	20	$120.00		$2,400.00

Are the labor rates detailed below fully loaded? **NO**

Provide any additional information and cost support data related to the nature of the direct labor detailed above.
Most of the research would be completed within the first 6 months. The software developer needs to provide additional documentation if required.

Direct Labor Cost ($): $2,400.00

Sum of all Direct Labor Costs is($):	$135,000.00

Overhead

Base

Labor Cost Overhead Cost	**$0.00**
Apply Overhead to Direct Equipment Cost?	**NO**
Overhead Comments:	
Overhead Cost ($):	**$0.00**

Year2

Labor Cost Overhead Rate (%)	**5**
Apply Overhead to Direct Equipment Cost?	**NO**
Overhead Comments:	
Overhead Cost ($):	**$120.00**
Sum of all Overhead Costs is ($):	**$120.00**

General and Administration Cost

Base

G&A Rate (%):	**1**
Apply G&A Rate to Overhead Costs?	**NO**
Apply G&A Rate to Direct Labor Costs?	**NO**
Apply G&A Rate to ODC- Equipment?	**NO**
Please specify the different cost sources below from which your company's General and Administrative costs are calculated.	
G&A Cost ($):	**$0.00**

Year2

G&A Rate (%):	**1**

Apply G&A Rate to Overhead Costs? **NO**

Apply G&A Rate to Direct Labor Costs? **NO**

Apply G&A Rate to ODC- Equipment? **NO**

Please specify the different cost sources below from which your company's General and Administrative costs are calculated.

G&A Cost ($): **$0.00**

Sum of all G&A Costs is ($): **$0.00**

ODC-Equipment

Base

Description: Neo4J

Vendor: Neo4J Inc

Quantity: 6

Total Cost ($): $876.00

Competitively Sourced? yes

Exclusive for this Contract? yes

Supporting Comments: **Neo4J license is required.**

Year2

Description: Neo4J

Vendor: Neo4J

Quantity: 1

Total Cost ($): $0.00

Competitively Sourced? yes

Exclusive for this Contract? no

Supporting Comments: **N/a**

ODC-Summary

Base

Do you have any additional information to provide? **NO**

Year2

Do you have any additional information to provide? **NO**

Profit Rate/Cost Sharing

Base

Cost Sharing ($): **-$0.00**

Cost Sharing Explanation:

Profit Rate (%):	**0**
Profit Explanation:	
Total Profit Cost ($):	**$0.00**

Year2

Cost Sharing ($):	**-$0.00**
Cost Sharing Explanation:	
Profit Rate (%):	**0**
Profit Explanation:	
Total Profit Cost ($):	**$0.00**
Total Proposed Amount ($):	**$135,996.00**

CERTIFICATE OF COMPLETION

THIS CERTIFICATE IS PRESENTED TO

Anand Thiagarajan, Datelite LLC

FOR SUCCESSFULLY COMPLETING FRAUD, WASTE AND ABUSE TRAINING AND MEETING ALL REQUIREMENTS SET FORTH BY THE OFFICE OF SMALL BUSINESS PROGRAMS

DEPARTMENT OF DEFENSE
UNITED STATES OF AMERICA

Nov 06, 2024	Nov 06, 2025
COMPLETION DATE	EXPIRATION DATE