Enhanced tracking and reporting of missing persons using Knowledge Graph and Ontology Engineering

Suparno Roy Chowdhury, Ayush Desai, Mrunal Kapure,
Mustakim Shikalgar, Ramchander Venugopal, Srividya Bansal
School of Computing and Augmented Intelligence
Arizona State University, AZ, USA
{srchowd3, adesai63, mkapure, mshikalg, rvenugo7, skbansa2}@asu.edu

***Abstract*—The issue of missing persons remains a significant concern in the United States, with thousands of cases reported annually. Addressing this challenge requires innovative methods to integrate and analyze disparate data sources to uncover patterns in disappearances. This research explores the potential of Ontology Engineering and Knowledge Graphs to provide a structured and interconnected perspective on missing person data. Ontologies enable the representation and linking of real-world entities within a unified framework, facilitating more meaningful data relationships.**

In this study, a knowledge graph is constructed to capture key details about missing persons, including the circumstances of their disappearance, personal characteristics, and last known locations. Additionally, a Large Language Model (LLM) is integrated to summarize query results, enhancing data interpretation. This knowledge graph serves as a foundation for advanced data analysis, enabling real-time insights, supporting proactive interventions, and aiding in case resolution.

***Index Terms*—Semantic web, Ontologies, Knowledge Graphs, Data Integration, Missing people, Risk factors, NamUS, LLM.**

I. Introduction

This study defines what constitutes a missing person, explores the causes behind disappearances, and examines how semantic web models construct knowledge graphs that connect relevant data. By developing ontologies around missing persons and related contextual information, this research creates a unified knowledge graph that integrates various data sources. This knowledge graph serves as the foundation for a human-centered intelligent system designed to enhance the analysis of missing-person cases. Additionally, by leveraging generative AI for advanced text processing and summarization, the system provides meaningful insights into missing person cases in California.

Howlings et al. [1] defines a missing person as an individual who cannot be located or whose well-being cannot be confirmed. The diverse causes of disappearances suggest that no single factor is solely responsible; instead, multiple risk pathways contribute to a person going missing [9]. This underscores the need for a decision-making system that integrates these risk factors with existing missing persons databases to facilitate search and rescue efforts.

A semantic web model, in the form of an ontology, captures information about this domain. Ontologies define the semantics of domain concepts, using structured terms to represent real-world objects [4]. They create a shared understanding of a domain, enabling consistent communication between people and software. By organizing entities (or classes) and their relationships in a hierarchical structure, ontologies allow subclasses to inherit properties from their superclasses. Additionally, they enable developers to define domain-specific relationships, such as whole-part associations or causal links, to model interactions between entities.

Ontologies are valuable for standardizing information, aligning perspectives across teams, supporting tool integration, and enabling automated reasoning. They provide a common language across systems, ensuring data consistency and facilitating seamless data validation. In complex and interdisciplinary fields such as safety analysis, ontologies enhance collaboration by creating a unified framework for knowledge representation [3].

Knowledge graphs (KG), along with their formal schemas - ontologies - facilitate data integration and population, enabling connections and the discovery of new insights [15]. According to the World Population Review, the United States has the highest number of reported missing persons cases worldwide. Given its robust reporting system and large population, this study constructs a knowledge graph to track these reports and generate insights into missing-person cases. The knowledge graph incorporates key factors such as age, gender, race, and reasons for disappearance.

Our approach employs semantic data representation using the subject-predicate-object triple format, following RDF and TTL standards [16]. Additionally, we integrate a Large Language Model (LLM) for query result summarization, demonstrating human-AI collaboration in providing real-time insights and proactive interventions for locating missing persons. The choice of data representation standards aligns with their compatibility with LLMs [18] and adheres to the philosophy of Linked Open Data [19].

The remainder of this paper is organized as follows: Section II discusses related work. Section III presents our proposed approach and high-level system design, followed by implementation details in Section IV. Section V evaluates our work, while Section VI discusses findings, challenges, and future directions. Finally, Section VII presents our conclusions.

II. Related Work

As this research aims to uncover patterns in why individuals go missing, it is essential to first identify potential categories that provide direction in recognizing these patterns. Howlings et al. [1], in their study on police use of social media for investigating missing children, outline various reasons for disappearances. These include escaping abusive living conditions, fleeing stressful environments, and dealing with physical or mental health issues [1].

Henderson et al. [2], an Australian-based research team, conducted surveys to identify additional reasons why people go missing. The survey results categorize these causes into: unintentionally wandering off or getting lost, running away to escape consequences, rebelling against authoritative figures (often parents), and safety concerns, such as kidnapping or harm [2]. These categories can be linked to the victims in our study to extract new insights from the knowledge graphs.

Yahya et al. [4] demonstrate the effectiveness of Knowledge Graphs in real-world applications. Their research on Knowledge Graphs for Industry 4.0 (I4.0) explored the development of a reference generalized ontological model (RGOM), which they showed "can facilitate a range of I4.0 concepts including improved asset monitoring, production enhancement, resource reconfiguration, process optimization, product orders and deliveries, and the product life cycle" [4]. This highlights the power of knowledge graphs in capturing intricate details of real-life objects.

For our study, we gathered data from NamUS, a government-run resource center for missing, unidentified, and unclaimed persons in the United States [11]. NamUs provides detailed information on missing persons reported by each state. We focused on three states—California, Texas, and Alaska—since the World Population Review states that California, Texas, and Florida have the highest number of open missing person cases [7]. Although Florida is the state with the third most number of cases, we chose Alaska due to its higher rate of missing persons per 100,000 people, as reported by the World Population Report [7].

After further deliberation, the research team decided to focus on California, the state with the highest number of missing persons reported annually. According to the World Population Review, California leads the United States in missing person reports, with 3,010 individuals reported missing in 2024. This provides a rich dataset for constructing ontologies, where each individual can be treated as a real-world object to build the knowledge graph around.

Additionally, the choice of California is supported by a study conducted by Johnson et al. [6], which highlights the state's diverse population. According to a survey by the Public Policy Institute of California, no single race or ethnic group constitutes a majority of the population [6]. This diversity proves advantageous, as it provides a broader dataset that can be linked to our use case of categorizing missing persons based on ethnicity.

By constructing this framework, we aim to uncover potential reasons why people in California go missing. Gomes Jr. et al. [5] developed a framework to analyze missing persons in Brazil, discovering that age plays a significant role in disappearances. They identified several factors related to age, such as slave labor, chemical dependency, debt, and memory loss. This demonstrates the value of constructing frameworks to gather new insights. This approach aligns with the findings of Howlings et al. [1], who identified physical and mental health issues as contributing factors to disappearances, and Henderson et al. [2], who categorized escaping adverse conditions, such as financial difficulties, as a key reason for missing persons.

Szekely et al. [10] developed DIG, a knowledge graph system aimed at combating human trafficking by using semantic technologies to establish connections with trafficking victims. Similarly, missing persons reports can be linked to safety concerns, as emphasized by Henderson et al. [2], further validating the approach of previous research teams in utilizing semantic web models to derive insights from missing persons data.

The goal of this research is to engineer a comprehensive ontology that can be extended and integrated with other related ontologies, facilitating the construction of a knowledge graph for missing persons. This knowledge graph can also be connected to other relevant knowledge graphs, such as those related to human trafficking, to generate new insights.

III. Proposed Approach and High-Level System Design

Our approach begins by retrieving data on all current missing persons cases in the state of California. The NamUS website provides a downloadable CSV file containing information on various attributes related to each victim, such as the case number, date of last contact (DLC), name, age at the time of disappearance, city, county, state, biological sex, and race/ethnicity.

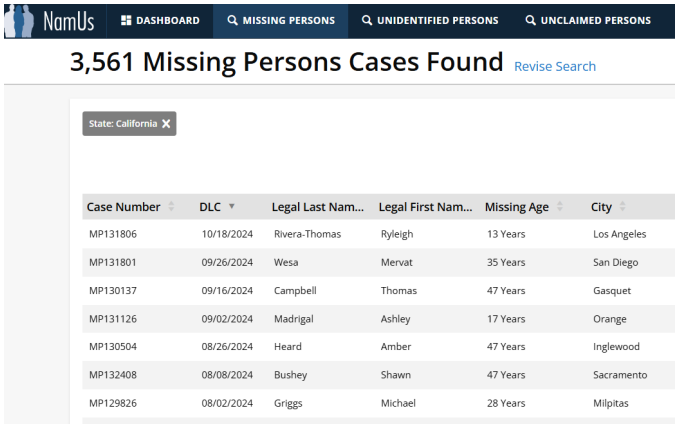

Fig. 1. The NamUS dashboard displaying data for California.

The website allows users to filter data by state, as shown in Figure 1, which significantly reduces the amount of scraping required and enables targeted data collection for specific regions. The use cases for our study are defined as follows:

- **Age-based**: Results categorized by the victim's life stage, e.g., Toddler, Teen, or Young Adult.
- **Location-based**: Queries based on the victim's last known location, specifically their city and county.
- **Race-based**: Queries based on the race or ethnicity of the victim.
- **Circumstance-based**: Queries related to the circumstances surrounding the victim's disappearance.

Once data on missing persons reports from California were retrieved via the CSV, we proceeded to scrape dynamically generated data from the website. The NamUS website makes API calls to display additional details for each victim (as shown in Figure 2). This information, which is not included in the downloadable CSV file, includes critical data such as the circumstances of the disappearance, current age, location of last contact, and physical description. However, since the API is not publicly accessible and these details fall under government domain, traditional scraping tools like BeautifulSoup could not be used. This information is crucial for our study, as it enables us to classify victims based on common themes or circumstances.

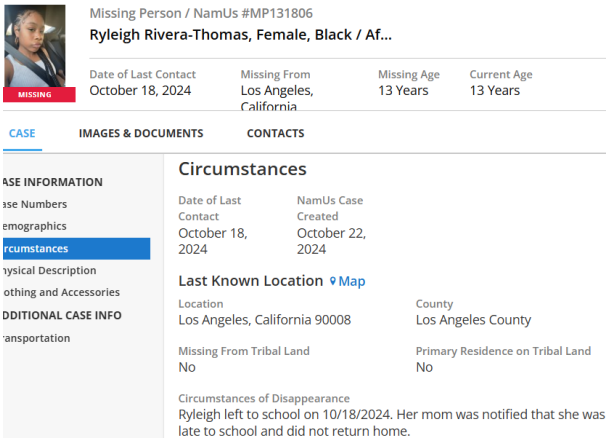

Fig. 2. A typical NamUS profile containing dynamic data not available in the CSV, such as the circumstances of disappearance.

To address this challenge, we used Selenium, a browser automation tool, to load individual profiles and retrieve the data generated dynamically by the API. Additionally, to pinpoint the last known locations of missing persons, we included a link to Google Maps, which provides quick access to their last point of contact. This link, along with the corresponding latitude and longitude coordinates, was extracted and added to the CSV file. This approach not only facilitates the visualization of each case but also supports the clustering of similar disappearance cases based on geographic location.

Following the retrieval of missing persons data for California, the next step was to clean and preprocess the data to make it more relevant to our use cases. The first issue encountered was the presentation of ages in full phrases, such as "13 Years" and "21 Years," rather than as integers. This format applied to both the "Missing Age" (the age at which the victim went missing) and the "Current Age" (the age the victim would be today if found). To address this, Python was used to convert these ages into integers, making data retrieval based on age categories more efficient.

For victims younger than a year at the time of their disappearance, the ages were displayed as "$<$ 1 Year" in the CSV file. While obtaining the exact date of birth for these victims and calculating their exact age proved difficult, the decision was made to round their ages to 0. This allowed them to remain categorized with victims under one year old.

The next preprocessing step involved merging the names. The scraped data contained separate columns for "Legal Last Name" and "Legal First Name." To prevent mismatches and simplify alphabetical sorting, the names were combined into a single column in the final output. Additionally, merging the names eliminated the risk of errors arising from common last names and ensured that name-based pattern matching would not interfere with the analysis.

TABLE I
FINAL LIST OF ATTRIBUTES FROM THE PROCESSED DATASET THAT ARE INCORPORATED IN THE ONTOLOGY

Relevant Attribute	**Description of Attribute**
Case Number	The unique identifier attached to a case
DLC (Date of Last Contact)	The last time the victim was contacted before they went missing
Legal Name	The legal name of the victim
Missing Age	The age that the victim went missing
City	City of origin for the victim
County	County of Origin for the victim
Biological Sex	The Biological Sex of the victim
Race / Ethnicity	The race and ethnicity of the victim
Current Age	The current age of the victim at the present date
Map Data	Google Map location of where the victim was last seen.
Latitude, Longitude	Coordinates of the place the victims were last seen
Circumstances of Disappearance	The last reported circumstances of the victim before they went missing.
URL	The hyperlink to the victim's NamUS profile and related case
Image URL	The hyperlink to display the victim's picture.

Next, we aimed to include the URLs for each individual case as a potential Uniform Resource Identifier (URI). The case numbers map to individual URLs, structured as follows: https://www.namus.gov/MissingPersons/ Case#/[*case number for the victim*]?nav. In the scraped data, case numbers are presented in the format "MP [*case number for the victim*]." During preprocessing, we combine these strings appropriately to form the correct URLs, which were then added as a new column in the dataset. The complete list of attributes used in the study is shown in Table I.

The scraping also retrieved information about the victim's physical description, including distinctive characteristics such as tattoos and scars. Since this information was unlikely to contribute to uncovering meaningful patterns and could

complicate the knowledge graph, it was excluded from the cleaned data.

Additionally, the scraped data included a 'date modified' field, which indicated the last update for each case. Given the current limitations of our platform in dynamically updating the knowledge graph for individual cases, this field was not included in the cleaned data.

Finally, the state column was removed. Since the dataset already focused on California, including the state column was redundant and did not add value to the analysis.

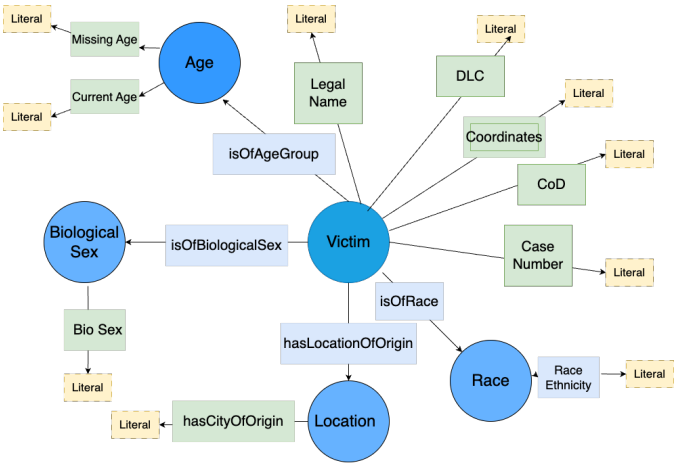

Fig. 3. Visualization of the missing persons Ontology

The missing persons ontology is a comprehensive system designed to organize and represent information about victims who have gone missing. Figure 3 provides a visualization of this ontology. It includes categories such as race (e.g. Asian, Black / African American, Caucasian) and crucial details like case numbers registered with NamUS, the locations from which individuals went missing, and the circumstances surrounding their disappearances. In addition, it tracks other relevant attributes of the victims, such as age, gender and location. The ontology also captures specific information, such as the last time the person was contacted, their legal name, their current age, and their age at the time of disappearance.

The user interface is implemented as a web application, powered by the knowledge graph. This platform allows users to perform queries based on the categories and attributes defined in the knowledge graph. For example, if a user wishes to analyze trends among victims based on race, the website offers the functionality to query this specific attribute. Similarly, users can query the knowledge graph based on other individual or combined attributes to extract insights.

The overall design and implementation of this intelligent web application system involved the following tasks:

- Fetch and pre-process data for populating the knowledge graph (KG).
- Construct the knowledge graph using the processed data in compliance with the ontology schema, represented in RDF triple format.
- Implement queries using SPARQL to fetch the required data after the knowledge graph is constructed.
- Build the web application that allows users to input their queries and view the results, while uncovering patterns related to missing persons.

IV. Implementation

Our implementation process begins with the development of our knowledge graph (KG) using the GraphDB database. GraphDB is a tool designed to create semantic graph databases that enable users to query data based on the relationships defined in the ontology [12]. These queries rely on the generation of a Terse RDF Triple Language (TTL) file [16], [18]. The TTL file establishes the relationships between the subjects, predicates, and objects, linking the data points from the CSV file and containing our pre-processed data.

In addition to the TTL file, our GraphDB instance requires the base OWL file, which contains the schematic details of our ontology for missing persons, including the classes, object types, and data types. The GraphDB library is hosted on a Microsoft Azure virtual machine, which is accessed by the web application used to query and present the data of missing persons in the state of California.

In our ontology design, we made a strategic decision to connect the city object type to a higher-level ontological concept—specifically, the DBPedia entry for city. This integration differentiates our database from NamUS by leveraging established ontologies to enrich our data and promote interconnectivity. By aligning our city object type with DBpedia, we enable our knowledge graph (KG) to link to a broader, well-established information repository, which enhances the richness and comprehensiveness of our datasets.

This approach demonstrates the potential for implementing interoperability between ontologies, a cornerstone of linked open data and a key strength of semantic web engineering [19]. By incorporating data from established ontologies, we not only enrich our knowledge graph but also increase its utility within the broader data ecosystem. For example, information about a missing person's city of origin can reveal patterns through the DBPedia ontology, providing deeper insights and connections that might help explain why they went missing in that particular city.

After constructing the knowledge graph, which serves as an actionable foundation for data retrieval, SPARQL queries were developed to extract relevant information regarding the victims. The system uses GraphDB as the backbone for constructing the knowledge graph on missing persons. SPARQL [17], an RDF query language, allows users to retrieve data based on specific use cases, such as individual names, case numbers, age, gender, race, location, and circumstances of disappearance. The following subsections discuss various SPARQL queries used for data retrieval. Each query serves a specific purpose, demonstrating how SPARQL facilitates structured and meaningful access to RDF data. These examples illustrate key concepts such as pattern matching, filtering,

and aggregation, showcasing SPARQL's versatility in querying semantic data stores.

A. Gender-based cases

For gender-based use cases, we developed queries based on the biological sex of the victim who had gone missing. This query can be used to study and determine which gender is at risk of missing in California and to formulate the reasons why they go missing, by analyzing their circumstances of disappearance. We have limited it to two, but are open to including more.

```
SELECT *
WHERE {
    //
     Write the triples
     set to retrieve relevant
     information about a victim.
    //
    rdfs:Bio_Sex ?sex;
    mpr:hasImageURL ?image
    FILTER REGEX(STR(?sex),
    [Male or Female])}
```

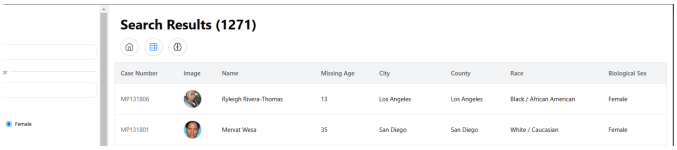

Fig. 4. The results of searching for all females

B. Age based cases

TABLE II
MISSING AGE RANGE

Life Stage	Age Bracket
Infant	0 - 12 Months
Toddler	1 - 5 Years
Kid	6 - 12 Years
Teen	13 - 17 Years
Young Adult	18 - 35 Years
Middle Aged Adult	36 - 55 Years
Senior	55+ Years

The query to return victims based on age ranges required filtering based on bounded values. The goal of this use case is to determine a pattern of disappearances based on age brackets. The age bracket shown would indicate the potential reasons why a victim within that age bracket might have gone missing. Furthermore, we would make this query based on missing age rather than current age, as the age the victim went missing would be more relevant to their situation, rather than querying for a cold case.

```
SELECT *
WHERE {
    ?victim a mpr:Victim;
    //
     Write the triples
     set to retrieve relevant
     information about a victim.
    //
    rdfs:isOfAge ?age;
    BIND(xsd:integer
    (REPLACE
    (STR(?missing_age), "^.*\#", ""))
    AS ?filter_age)
    FILTER ((?filter_age >= [minimum age]
    && ?filter_age <= [maximum age]) )
}
```

C. Location-based cases

Given that our tool is querying across one state, it significantly cuts down the amount of work needed to query based on location. However, it is still important to consider the counties and various cities in California. A certain city in a county could potentially exhibit a higher amount of people going missing, or the nature of this location might make it more prone for a victim to go missing.

```
SELECT *
    WHERE {
    ?victim a mpr:Victim;
    // Write the triples set to
    retrieve relevant
    information about a victim.
    //
    rdfs:hasCityOfOrigin
       ?city_of_origin;
    rdfs:hasCountyOfOrigin
        ?county_of_origin;
    rdfs:Coordinates ?cords;
    FILTER REGEX(
    STR(?city_of_origin),
      "[Insert City]"
    )
    FILTER REGEX(
    STR(?county_of_origin),
     "[Insert County]"
    )
}
```

Note: Not all cities belong to a specific county, and not all counties contain every city within their boundaries. Users can request data for a county but cannot individually check for a city. The front-end will display all relevant cities in that county. This design choice ensures that users are not confused as to why a city-county combination is not yielding results and prevents them from being overwhelmed by the massive list of cities across California. The users will need to know the county of the city they wish to search.

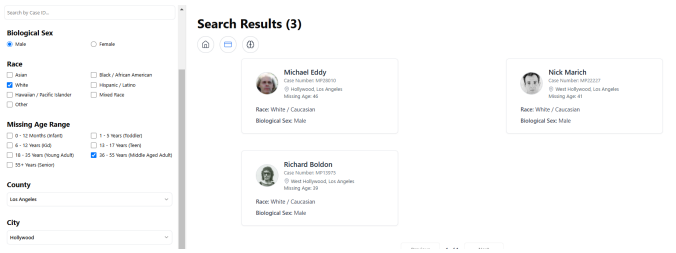

Fig. 5. Results of White males between 36 to 55 that have gone missing in LA County, Hollywood.

Furthermore, our counties are represented as objects rather than literals. For certain counties, the query will require inserting %20 to represent spaces. For example, ”Los Angeles” should be written as ”`Los%20Angeles`” in the query.

D. Race-based cases

Given California’s rich cultural diversity, it is important to recognize that no single race or ethnic group constitutes a majority of the population, as highlighted in the study by Johnson et al. [6].

In the context of our query, we considered the following racial and ethnic categories:

Race / Ethnicity:

- Asian
- White
- Hawaiian / Pacific Islander
- Other
- Black / African American
- Hispanic / Latino
- Mixed Race

To filter the dataset based on race, the following query setup can be used. It is crucial that the selected race is properly enclosed within the word boundary marker to ensure that only whole-word matches are returned. This is needed to avoid confusing entries like Asian and Caucasian together but at the same time, allowing mixed-race victims to show up in the result.

```
SELECT * WHERE {
    ?victim a mpr:Victim;
    // Write the triples set to
    retrieve relevant
    information about a victim.
    //
    rdfs:Race_Ethnicity ?race;
    FILTER (
    REGEX (
    LCASE(STR(?race)),
    "\b[Insert a race]\b" <--- boundary
    )
  )
}
```

Much like the age filter, multiple races can be entered by using the SPARQL ’or’ operator.

```
SELECT *
    WHERE {
    ?victim a mpr:Victim;
    // Write the triples set to
    retrieve relevant
    information about a victim.
    //
    rdfs:Race_Ethnicity ?race;
    FILTER
    (REGEX (LCASE(STR(?race)),
    "\\b[Insert Race]\\")
    ||
    REGEX (LCASE(STR(?race)),
    "\\[Insert Race]\\b"))
 }
```

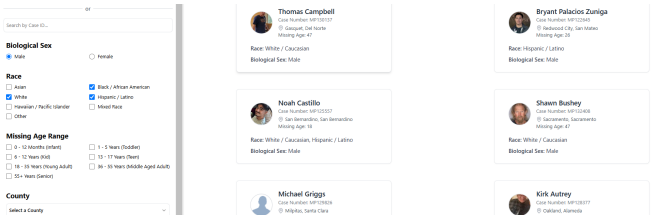

Fig. 6. Result of query looking for White, Hispanic, and African American males that have gone missing

E. Circumstance based query

NamUS provides reports on the circumstances surrounding a victim’s disappearance. The quality of these reports varies, with some offering detailed accounts while others may be incomplete or absent altogether. Despite this inconsistency, the reported circumstances remain valuable for analysis, particularly when examined in conjunction with other relevant categories, to gain insights into the potential reasons behind a victim’s disappearance.

After processing the data, we employed NLTK [20], a natural language processing tool, to analyze trigram combinations within the reported circumstances of disappearance. Trigrams, which are sequences of three consecutive words, help identify recurring patterns and phrases linked to the reasons individuals go missing. By examining these trigram combinations, we can uncover common themes and gain valuable insights into the typical circumstances that contribute to disappearances.

TABLE III
TOP 3 CIRCUMSTANCES BASED ON TRIGRAM RELEVANCY

Circumstance	Recurrence
(’foul’, ’play’, ’suspected’)	40
(’allegedly’, ’abducted’, ’mother’)	24
(’allegedly’, ’abducted’, ’father’)	22

The query to check for abduction cases is as follows.

```
SELECT *
    WHERE {
        ?victim a mpr:Victim;
        // Write the triples set to
        retrieve relevant
        information about a victim.
        //
        mpr:hasCircumstanceOfDisappearance ?cod;
        FILTER REGEX(
         LCASE(STR(?cod)),
         LCASE("[List circumstance]"), "i")
    }
```

We make sure to set both the circumstance of disappearance as lowercase to let the results match. We also provide the user the option to put in their own custom circumstances that they may be interested in studying within the California dataset. This can be seen in figure 8 which shows a number of missing-person cases that involved runaways.

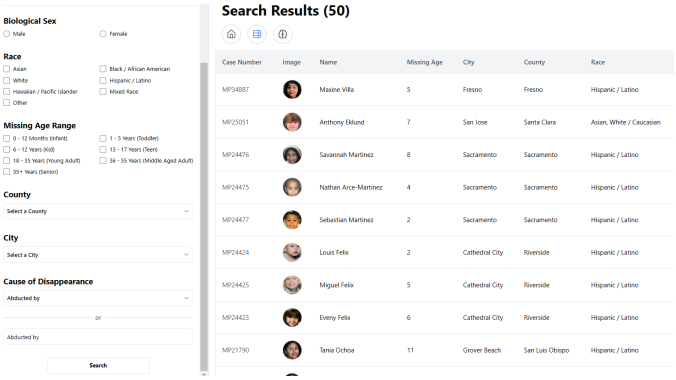

Fig. 7. Abduction cases

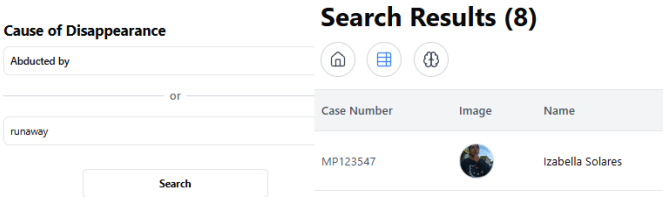

Fig. 8. Custom case with runaways. Runaway is entered in the search bar and retrieves all the runaway cases.

F. Large Language Models

In addition to querying for information, our toolset integrates LLM capabilities to provide insights based on the results obtained from the SPARQL queries. We have integrated OpenAI's GPT-3.5 Turbo Model to summarize the information returned from these queries [13]. These models excel at processing natural language inputs in a conversational style, generating meaningful outputs based on the prompts provided. According to a study by Gao et al., "ChatGPT has the ability to perform summarization evaluation using various human evaluation methods" [14], though the effectiveness of the summarization relies heavily on crafting a well-structured prompt to clearly convey the relevant information.

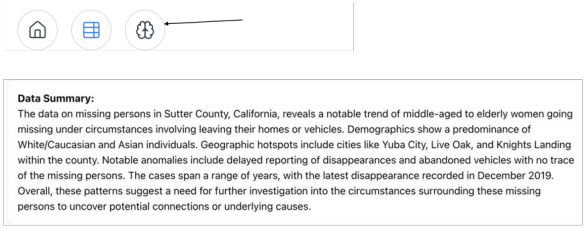

Fig. 9. Prompt result when attempting to analyze filtered results of missing female victims in Sutter County.

The GPT-3.5 Turbo Model is accessed through OpenAI's API, which allows us to pass victim data as a JSON object. The model is then prompted to identify and highlight interesting patterns and anomalies within the data retrieved from our knowledge graph. The prompt used for our proposed system is as follows:

"Summarize the following data on missing persons into a single concise paragraph, focusing on overall patterns, trends, and notable anomalies across the cases. Avoid listing individual case names and instead highlight key insights such as recurring demographics, geographic hotspots, timeframes, or shared circumstances. The goal is to provide actionable observations for researchers studying missing persons. Use professional and precise language, avoiding redundancy or vague phrases."

Figure 9 shows the results of using the prompt for a specific query of missing female victims in Sutter County.

V. Evaluation

To assess the effectiveness of our study, we conducted a comparative analysis with the NamUS website, which served as the source of our dataset. Utilizing NamUS as a baseline allows us to evaluate the enhancements and features integrated into our tool, highlighting the distinct advantages and improvements our system offers over the existing capabilities provided by NamUS.

TABLE IV
Evaluation: Feature Comparison of Proposed KG System with NamUS

Feature	Proposed ontology-driven KG System	NamUS
GPT-3.5 Integration	Integrated GPT-3.5 to parse large textual data, provide succinct summaries, and identify patterns.	No AI integration for text analysis or summarization.
Visualization Options	Includes visualization tools like pie charts (e.g., Figure 12) to help users draw actionable insights.	No visualization features available.
Category Selection for Circum-stances	Offers pre-selected categories based on recurring circumstances (e.g In Table III), simplifying classification.	Requires manual input for categorization as seen in Figure 10
Automation in Pattern Recognition	Automatically identifies patterns and trends within the data through AI capabilities.	Manual analysis required to detect patterns or trends.
Insights from Historical Data	Provides insights by detecting recurring patterns in historical cases.	Does not leverage historical data insights effectively.

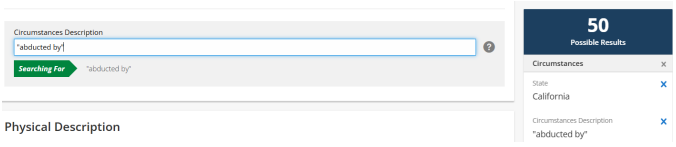

Fig. 10. The circumstance of disappearance search bar in NamUS. Our toolset offers pre-categorization and enables the user to enter their own categories.

During our experimentation with the tool, GPT-3.5 identified an intriguing anomaly involving instances of abandoned cars. By manually entering the phrase "found abandoned" into the search bar, we uncovered additional

relevant results for further analysis. Expanding the search to include cases of victims disappearing near the Golden Gate Bridge revealed a previously unnoticed pattern, opening a new avenue for exploration. Figure 11 illustrates this thought process in action.

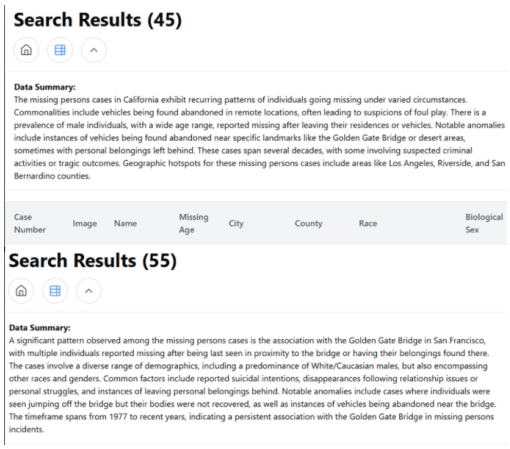

Fig. 11. Abandoned Vehicles found near Golden Gate reveals a new pattern

The summarization capabilities of GPT-3.5 significantly reduce investigation time and empower users to independently uncover valuable insights. In addition to the advanced summarization capabilities offered by GPT 3.5, our toolset provides additional support with visualizers for the data. We extract relevant attributes on the victims from our SPARQL queries to create pie charts as seen in Figure 12. Detailed SPARQL queries are not included due to page limit constraint.

Given the novelty of the field of LLMs, we are aware that our usage of this tool in the context of drawing insights needs formal investigation. We will continue to explore this idea as part of our future work.

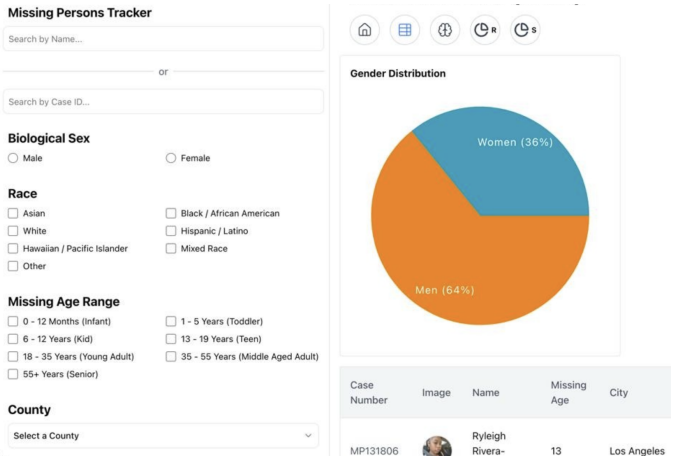

Fig. 12. Pie chart returning the distribution of missing victims based on gender

VI. Discussion, Challenges, and Future Work

In this section, we discuss some of the challenges with the construction of the proposed KG. Formulating SPARQL queries based on race proved challenging. Problems ranged from not being able to accurately query for mixed race victims to returning results for victims with similarly spelled races (i.e. Asian and Caucasian). Querying would sometimes return incorrect data, and although fixes have been implemented to address these particular shortcomings, more testing will be needed to discover edge cases. Furthermore, GPT 3.5 has a limited number of tokens it can process before it fails to generate a summary, with the upper limit coming at approximately 90 results. NamUS data also has inconsistencies within its records. For instance, one record may provide a detailed explanation of the circumstances leading up to a person going missing, while another offers only a vague explanation, as can be seen in Figure 13. This study aimed to build a prototype to test the efficacy of the ideas, and our initial evaluation showed promising results. The challenges listed above will be the focus of our future work.

Fig. 13. Difference between description in circumstances.

For future work, we also plan to expand this setup to include more states, such as Alaska and Texas among others. Currently, we have focused on California due to its diversity in cases and high density of missing persons per square area. We also aim to further integrate LLMs such as GPT 3.5 to aid in the construction and engineering of the knowledge graph. Studies by Shimizu, et al have shown that, by setting up conceptual boundaries, LLMs have the ability to tackle "a variety of tasks, seeing already success in ontology construction, (complex) alignment, and population" [15]. Noting this, we believe that with a more integrated approach with existing LLMs, we can establish rigid knowledge graphs with the ability to form solid profiles for each victim, allowing for detailed entities that enhance the data retrieval and investigation process.

VII. Conclusions

Semantic web models provide a unique and robust framework for constructing interconnected databases by leveraging subject-predicate-object triples. This approach

goes beyond the conventional design that merely retrieves information, offering the ability to contextualize and link the real-world entities and relationships. Our proposed missing person tracker tool utilizes this methodology and demonstrates significant potential to uncover insights into why individuals go missing. Categories of circumstances of disappearance have already been identified to some extent, and scaling the research could lead to the discovery of additional patterns. Furthermore, the knowledge graph is well-suited for linked open data initiatives with the integration of other ontologies [19]. When combined with the advanced summarization capabilities of large language models (LLMs), the ability to derive and interpret new insights is greatly enhanced, paving the way for deeper understandings and broader applications.

REFERENCES

[1] Eleanor Howlings, Reka Solymosi. Exploring facilitators, barriers and concerns of police using social media when investigating missing children. International Journal for missing persons.vol 1. Issue 1. Available : https://scholarworks.sjsu.edu/cgi/viewcontent.cgi?article=1003&context=-ijmp

[2] Monica Henderson, Peter Henderson. missing persons : Incidence, Issues and Impact. Available: https://www.researchgate.net/publication/260401235_Persons_-Incidence_Issues_and_Impacts

[3] Rafael Batres, Shinya Fujihara, Yukiyasu Shimada, Testuo Fuchino. "The use of ontologies for enhancing the use of accident information". Available : https://www.sciencedirect.com/science/article/abs/pii/S095758201200-1449

[4] Mohammad Yahya, John G. Breslin, Mohammad Intizar Ali. "Semantic Web and Knowledge Graphs for Industry 4.0". Available : https://www.mdpi.com/2076-3417/11/11/5110

[5] Jorão Gomes,Jr., Nicolas Ferranti, Jairo Francisco de Souza. Semantic Enrichment of Web Data for the Provision of an Unified Data Repository of Brazilian missing persons. Available : https://dl.acm.org/doi/10.1145/3330204.3330267

[6] Hans Johnson, Marisol Cuellar Mejia, Eric McGhee. "California's Population". Available : https://www.ppic.org/publication/californiaspopulation/#:̃text=No%20race-%20or%20ethnic%20group,the%202022%20American%20Community-%20Survey

[7] World Population Review. "Missing Person by State 2024". Available : https://worldpopulationreview.com/state-rankings/missing-persons-by-state

[8] World Population Review. "missing persons Statistics by Country 2024". Available : https://worldpopulationreview.com/country-rankings/missing-persons-statistics-by-country

[9] Lorna Ferguson. "Risk Factors and missing persons: advancing an understanding of 'risk'". Available : https://www.nature.com/articles/s41599-022-01113-8

[10] P. Szekely et al., "Building and Using a Knowledge Graph to Combat Human Trafficking.". Available: https://usc-isi-i2.github.io/papers/szekely15-iswc.pdf

[11] National Institute of Justice. (2024). National Missing and Unidentified Persons System (NamUs) [Dataset]. U.S. Department of Justice. https://www.namus.gov

[12] GraphDB downloads and resources.https://graphdb.ontotext.com/

[13] Open AI (GPT-3.5 Turbo). [Online]. Available: https://platform.openai.com/docs/models/gp#gpt-3-5-turbo

[14] M. Gao, J. Ruan, R. Sun, X. Yin, S. Yang, and X. Wan, "Human-like Summarization Evaluation with ChatGPT," arXiv.org, Apr. 05, 2023. https://arxiv.org/abs/2304.02554

[15] C. Shimizu and P. Hitzler, "Accelerating Knowledge Graph and Ontology Engineering with Large Language Models," arXiv.org, 2024. https://arxiv.org/abs/2411.09601.

[16] W3C RDF Working Group. "RDF 1.2 Concepts and Abstract Syntax," W3C Recommendation. 2023. Available at: https://www.w3.org/TR/rdf12-concepts/.

[17] W3C RDF-star Working Group, "SPARQL 1.2 Query Language," W3C Recommendation, 2023. [Online]. Available: https://www.w3.org/TR/sparql12-query/.

[18] J. Frey, G. Ladwig, and T. Liebig, "Benchmarking the Abilities of Large Language Models for RDF Knowledge Graph Creation and Comprehension: How Well Do LLMs Speak Turtle?," arXiv preprint, arXiv:2309.17122, Sep. 2023. [Online]. Available: https://arxiv.org/abs/2309.17122.

[19] T. Berners-Lee, J. Hollenbach, and M. Fischetti, "Linked Open Data: The Philosophy and Practice of Making Data Accessible," IEEE Internet Computing, vol. 15, no. 3, pp. 16-23, 2011.

[20] E. Loper and S. Bird, "NLTK: The Natural Language Toolkit," Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching NLP and CL, 2002, pp. 63–70.