NYC CrimeWatch: Crime Analysis Tool

Subham Kumar
1224913131
skuma251@asu.edu

Amarjeet Singh
1226070601
asing499@asu.edu

Heet Punjawat
1225621982
hpunjawa@asu.edu

Paromita Roy
1224708511
proy13@asu.edu

Siri Jarmale
1225221998
sjarmale@asu.edu

***Abstract*—This project aims to develop a data-driven web application that analyses crime-related data in New York City. It will specifically target shooting incidents, and hate crimes, with data related to their complaints and arrests. The application will utilize ontologies to organize and contextualize the data, providing valuable insights for informed decision-making and crime prevention. This initiative will focus on using Semantic Web Engineering to make use of raw data from the web and transform it into a meaningful manner to provide insights.**

Keywords - Hate crime, shooting, ontologies, RDF, Semantic Web.

I. Introduction

The growing concerns related to public safety and law enforcement in the urban environment of New York City have prompted the need for a comprehensive and data-driven approach to understand and address crime-related issues. While the city thrives on its diversity and multiculturalism, it is essential to recognize the unfortunate reality that hate crimes persist. These crimes, driven by prejudice and intolerance, manifest in various forms, including vandalism, harassment, and physical violence. One of the most alarming manifestations of this hatred is when it culminates in shooting incidents. In 2021, hate crime incidents in New York increased by 55% [1].

To tackle the complex and interconnected nature of crime data, we will employ techniques in Semantic Web Engineering [11]. We will leverage the power of ontologies and RDF (Resource Description Framework) to structure and semantically enrich raw data. This transformation will enable seamless connections between diverse information sources, facilitating a deeper understanding of crime patterns and trends. Converting data from various formats, including XML, into RDF will lay the foundation for comprehensive data analysis.

II. Problem Statement

The interconnected nature of mass shootings, hate crimes, and criminal complaints remains a matter of significant concern. This nexus, where violence, bias, and formal documentation of crimes converge, presents a complex and multifaceted problem. Various law enforcement agencies extensively cover these incidents, generating a substantial volume of data. This data typically includes details of the crimes, perpetrators, locations, and victims. Yet, the true value of this data in comprehending crime patterns and facilitating proactive measures remains underutilized.

To address this issue, our project aims to harness the power of ontologies – structured representations of knowledge – to organize and make sense of this crime data[3]. We seek to create a systematic structure that can better contextualize the data we collect, making it more comprehensible and actionable.

In response, we propose the development of an innovative application that aims to collect, analyze, and present a wide range of crime data, with a specific focus on shooting incidents, complaints, arrests, and hate crimes occurring in different boroughs of New York. By aggregating, processing, and visualizing data from various sources, the application will provide stakeholders with valuable tools for informed decision-making, resource allocation, and proactive crime prevention.

Use Cases:

Shooting Data by Race (Perpetrator and Victim) Over Time: The application will enable users to explore shooting incident data broken down by race for both perpetrators and victims over the years. This feature will provide insights into potential racial disparities in shooting incidents.

Temporal Analysis of Shooting Incidents: Users can access a feature that counts the number of shooting incidents per month and per year, providing a clear understanding of when these incidents tend to occur and whether there are any temporal patterns.

Ratio of Shooting Incidents Relative to Total Reported Crimes: The application will calculate and display the ratio of shooting incidents to the total number of reported crimes. This ratio offers a measure of the prevalence of shootings within the broader crime landscape.

Demographic Analysis of Shooting Incident Ratios: Users will have the ability to analyze shooting incident ratios relative to total crimes, considering factors such as race, age, or borough. This allows for a more nuanced examination of disparities and trends.

Top 5 Crimes by Gender Annually: The application will present a list of the top 5 crimes, specific to each year, for both genders. This information helps identify evolving crime trends and patterns related to gender.

Hate Crime Analysis: The application will provide insights into hate crimes, their frequency, and demographics. Users can analyze hate crime data by location, time, and other relevant factors to understand patterns and connections. Specifically, the application can help identify areas where hate crimes are most prevalent and break down hate crimes by race or ethnicity, facilitating efforts to combat hate crimes and promote tolerance and safety in affected communities.

These use cases, enabled by the application, will empower stakeholders with the knowledge needed to make informed decisions, allocate resources effectively, and implement proactive measures to address the interconnected nature of mass shootings, hate crimes, and criminal complaints in New York.

III. Literature Review

In the realm of crime analysis, mapping, and data-driven crime prevention, it is evident that innovative approaches are essential for comprehending and addressing various forms of crime, especially hate crimes and shooting incidents within complex urban environments. These incidents rarely occur in isolation; they are interconnected with other criminal activities, like hate crimes and criminal complaints, making traditional data analysis methods struggle to establish relationships and context effectively.

Although shootings in New York City saw a 20% decrease earlier this year [3], still it remains a significant concern. Understanding and analyzing these patterns offer the potential to reduce these statistics. Researchers emphasize the importance of utilizing advanced technologies like the Semantic Web, ontologies, and RDF to construct a coherent framework for comprehensive data analysis.

Driven by a strong dedication to data analysis and technology, the concerning increase in hate crimes and shootings has motivated us to embark on the development of a comprehensive analytical framework. The need to gather, interpret, and present data that sheds light on the root causes, evolving trends, and recurring patterns associated with these incidents is paramount. This analytical endeavor has the potential to empower law enforcement agencies, policymakers, and communities to make well-informed decisions, implement precise interventions, and contribute to the creation of a safer and more inclusive society. It is a resounding call to harness the power of data and technology in order to confront and mitigate the escalating threats posed by hate-motivated violence.

Historically, XML and JSON have been standard data formats for web communication. However, XML has limitations in achieving semantic interoperability and data comprehensibility due to varying metadata standards. In such cases, converting XML data into RDF is necessary. 'XML to RDF Conversion: A Generic Approach' [4] offers valuable insights for this purpose.

The next step in our project will be the design of an ontology to capture the relevant details to be used in our project. Ontologies: How can They be Built? [5], the paper proposed a skeletal methodology for ontology development. Also, it gives us detailed information about the process that needs to be followed while producing an ontology. Apart from this, different techniques have been mentioned to build the ontology. By looking at all the techniques, ENTERPRISE process looks more relevant to use in our use case. The proposed enterprise approach outlines four stages for ontology development: purpose identification, building the ontology, evaluation based on specific criteria, and documentation. The ontology development stage involves identifying the essential concepts and relationships in the domain, crafting clear and unambiguous definitions, and selecting appropriate terminology as per their specific problem statement.

Various papers have been published discussing the creation of specialized ontologies for the realm of criminal investigations. These ontologies, using RDF and OWL technologies, are designed to categorize and structure knowledge within the crime domain. Knowledge Engineering and Ontology for Crime Investigation [6] discusses the PREVISION ontology which is centered around the intelligent pentagram framework (widely used in the field of intelligence analysis). This includes vital concepts like events, places, persons, and equipment while building the ontology and can help in linking the relevant data with another incident.

As we all know visualization plays a vital role in managing complex data, such as ontologies. Visualization techniques help us to represent the data in the form of two-dimensional trees or graphs showing different concepts and relationships. Study done by M. Lanzenberger, J. Sampson and M. Rester [7] discusses the various tools that are available in the market, each using different visualization methods and offering different interpretations. Protege [8][9] is an open-source tool that can be used for visualizing the developed ontologies. The main advantage of using this tool is its flexible plug-in architecture which can accommodate the creation of ontology-based applications, ranging from straightforward to intricate ones.

The extracted information is used to construct a unified knowledge graph, typically represented as a set of triples with subject, predicate, and object, following RDF standards. To query data from multiple RDF graphs, SPARQL is a valuable tool. For an in-depth understanding of SPARQL, its syntax, and semantics, you can refer to the 'Querying Semantic Web Data with SPARQL' paper [10]. This paper also addresses some of the challenges encountered when working with linked data.

With all these steps, our application aims to show different visualizations of the different incidents that happened in New York State along with some additional details.

IV. Approach and High Level Design

Our project simplifies the process of transforming raw crime data from various sources, including Shooting Incident Data, Hate Crime data, and Complaints and Arrests Data in CSV format. After data processing, the processed data has to be fed into our OWL ontology model created with Protege using the Cellfie plugin. This ontology model is populated with real-world instances into an RDF file. To facilitate user queries, Apache Jena Fuseki is utilized as a standalone server, leveraging SPARQL for precise data filtering. We expose the functionality of Jena and Fuseki through RESTful APIs developed with Java Springboot, which act as the communicative conduit between the backend operations and the frontend interface. The user interface, crafted with Vue.js, presents an accessible and interactive platform for users to engage with and manipulate the data. By integrating

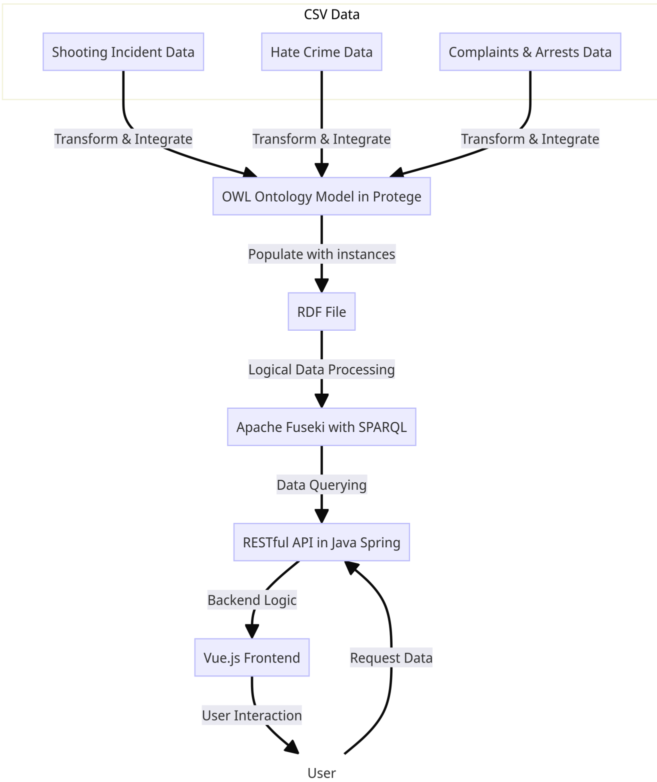

Fig. 1. High Level Design and Workflow

RESTful APIs, we ensure a seamless, scalable, and flexible interaction between the frontend and backend, adhering to our commitment to simplicity and efficiency while enabling a thorough exploration of NYC crime data.

V. Ontology Design and Visualization

The ontology serves as a structured framework for capturing essential details about crime incidents and their contextual information to meet our project's use cases. It includes major classes such as perpetrator and victim information, crime location, crime type, and the data source. Each crime incident is uniquely identified by a crime ID and includes the occurrence date and time.

To provide a detailed representation of crime locations, the Location class is further categorized into two subclasses: Boro and Precinct. Object properties, such as "has victim" and "has perpetrator," establish connections between the crime class and the victim and perpetrator classes, facilitating the association of individuals with specific incidents.

The various object properties are defined to link different classes, enabling a comprehensive understanding of crime incidents. Data properties, including age, sex, race, and precinct ID, are defined to relate instances to specific data values. This comprehensive ontology design underpins the project's data organization and semantic enrichment, enhancing the analysis of crime-related information. The ontology's structure and relationships can be visualized in Figure 2.

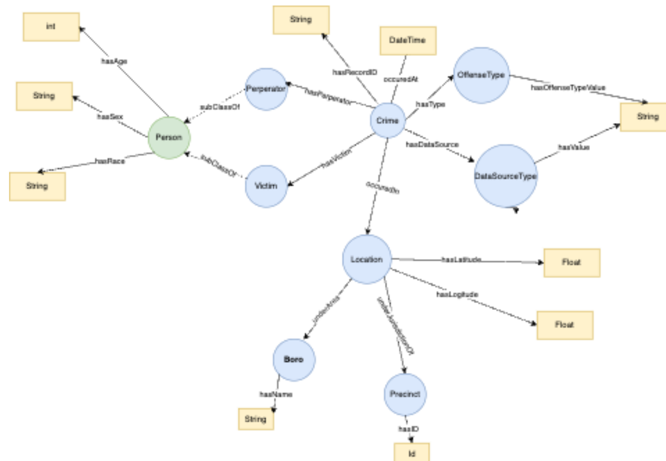

Fig. 2. Ontology Design

VI. Data Collection and Preprocessing

The dataset for our use case has been extracted from the New York City open data website [12]. This website provides different types of datasets related to different incidents happening in the city, offering valuable datasets for analysis and research purposes. Our primary focus has been on three different NYPD datasets related to shootings, events, and hate crimes as they align closely with our project's goals.

The first dataset, NYPD shooting incident data [13], provides all the information about the shooting incidents With 21 columns and 27.3K rows, each row in this dataset represents a unique shooting incident. There are various variables such as the "Occurrence date and time", Borough along with latitude and longitude values for geographical context, and other details containing information about the perpetrator and victim and much more. These columns are of great significance in our project.

The hate crime dataset [14] provides us the detailed information related to the different hate crimes that occurred in New York City. In this raw dataset, there are a total of 2.2k rows and 14 columns. These different columns provide essential conceptual information about each distinct hate crime incident. The dataset contains columns such as complaint ID, and different demographic details such as borough name and county name, Offense Type, and occurrence date and time which play an important role in our use cases. Some columns such as arrest date, arrest id, and duplicate columns such as month number and complaint year do not completely align with our objectives, so these are being dropped.

Our analysis is further enriched by the inclusion of the NYPD Complaint Data [15], which offers a comprehensive view of various types of incidents reported in New York City. This dataset boasts a substantial 35 columns and a vast 8.35 million rows, with each row corresponding to a distinct complaint. The dataset's focus aligns with essential attributes such as "Date and Time of Occurrence," "Borough of Occurrence" for geographical context, "Level of Offense" for gauging incident severity, and detailed information regarding ”Suspect Details” and ”Victim Details.” Additionally, ”Arrest Date” assumes paramount importance in understanding the resolution of complaints. The analysis of this dataset yields profound insights into the wide spectrum of complaints and incidents within the city.

To fulfill our use cases, data preprocessing is required. As all three datasets have missing and inconsistent values, proper cleaning of data is essential. Apart from this, some columns that are either duplicates or are of no use in our use cases have been dropped during this process. As there are few rows with missing column values, therefore the rows having required column values missing are dropped to maintain consistency. Also, the categorical values such as age groups are also handled by assigning some integer values to each group for the ease of querying the data and linking it with other datasets. Python script is used for this to avoid the manual error and reduce the time required. This script uses various Python libraries such as Pandas and Numpy to complete this task.

VII. ROLES AND RESPONSIBILITIES

Name	Responsibility
Amarjeet Singh	Responsible for ontology design, ensuring that it accurately represents crime-related data and aligns with project requirements.
Subham Kumar	Will lead the development of the user interface, ensuring a user-friendly and visually appealing frontend application.
Heet Punjawat	In charge of data cleaning and preparation, including removing irrelevant columns and handling missing or inconsistent data to make it suitable for RDF conversion. After data preprocessing, will collaborate with Paromita on the project.
Paromita Roy	Responsible for the development of the Spring Boot API, enabling SPARQL queries for data retrieval and ensuring well-structured endpoints with comprehensive documentation.
Siri Jarmale	Responsible for developing SPARQL queries, ensuring efficient data retrieval and analysis for the project.

VIII. IMPLEMENTATION PLAN AND TASKS TO BE COMPLETED

The implementation phase of our project is a well-organized process that involves a series of crucial steps which are discussed below:

Ontology Design: In the initial phase of implementation, the team will focus on ontology design using Protege. This process will include the definition of classes, properties, and relationships that accurately represent crime-related data. The ontology design is critical for providing a structured framework that makes data organization and semantic enrichment possible. The team will ensure that the ontology aligns with the project’s specific requirements and use cases.

Data Cleaning and Preparation: Following ontology creation, the project will move to the data cleaning and preparation phase. The raw datasets obtained from the New York City Open Data website will be processed to ensure data quality. Irrelevant columns will be removed, and strategies will be employed to handle missing or inconsistent data. This meticulous data preprocessing is essential to make the datasets suitable for RDF conversion.

RDF Conversion: The cleaned and refined data will then be converted into RDF format using the Cellifi plugin in Protege. This step is crucial as it transforms the tabular data into a linked data format that adheres to the ontology design. RDF triples will be generated to establish relationships between different entities, connecting various aspects of crime incidents, including victims, perpetrators, locations, and temporal information.

Triple Store Setup: To facilitate efficient data storage and retrieval, the RDF data will be hosted on Apache Fuseki, serving as a triple store. This setup allows for the organized storage and management of semantic data. With RDF data loaded into Fuseki, users will have the ability to query the data effectively and gain valuable insights.

Development of Spring Boot API: The project will include the development of a Spring Boot API that interfaces with the Fuseki triple store. This API plays a central role in enabling SPARQL queries for data retrieval. The team will ensure that the API offers well-structured endpoints and comprehensive documentation, making it user-friendly and accessible. Users will be guided through the query process to retrieve data aligned with predefined use cases.

Frontend Application: Simultaneously, a frontend web application will be created using modern web technologies and JavaScript frameworks. Data visualization libraries, such as D3.js and Chart.js, will be integrated to provide visual representations of the retrieved data. The application will be designed with a user-friendly interface, allowing users to interact with the data effortlessly. User interaction capabilities will be implemented to enable exploration of various predefined use cases. Visualizations and analytical results, including charts, graphs, and maps, will be presented within the frontend application, providing stakeholders with valuable insights for informed decision-making.

REFERENCES

[1] Hate crime in New York State 2021 annual report. (n.d.). https://www.criminaljustice.ny.gov/crimnet/ojsa/hate-crime-in-nys-2021-annual-report.pdf

[2] M. @ M. Abdul Jalil, C. P. Ling, N. M. Mohamad Noor, and F. Mohd., “Knowledge Representation Model for Crime Analysis,” \textit{Procedia Computer Science}, vol. 116, pp. 484–491, 2017, doi: https://doi.org/10.1016/j.procs.2017.10.067.

[3] H. Meko, “Shootings in New York drop by a quarter as surge of violence eases,” The New York Times, https://www.nytimes.com/2023/07/06/nyregion/shootings-nyc-crime.html (accessed Oct. 21, 2023).

[4] “XML to RDF Conversion: A Generic Approach,” ieeexplore.ieee.org. https://ieeexplore.ieee.org/abstract/document/4688061 (accessed Oct. 22, 2023).

[5] H. S. Pinto and J. P. Martins, “Ontologies: How can They be Built?,” *Knowledge and Information Systems*, vol. 6, no. 4, pp. 441–464, Jul. 2004, doi: https://doi.org/10.1007/s10115-003-0138-1.

[6] Wilmuth Müller, D. Mühlenberg, D. Pallmer, U. Zeltmann, C. Ellmauer, and Konstantinos Demestichas, “Knowledge Engineering and Ontology for Crime Investigation,” *Springer eBooks*, pp. 483–494, Jan. 2022, doi: https://doi.org/10.1007/978-3-031-08333-4_39.

[7] M. Lanzenberger, J. Sampson and M. Rester, ”Visualization in Ontology Tools,” 2009 International Conference on Complex, Intelligent and Software Intensive Systems, Fukuoka, Japan, 2009, pp. 705-711, doi: 10.1109/CISIS.2009.178.

[8] Stanford Center for Biomedical Informatics Research. (n.d.). A free, open-source ontology editor and framework for Building Intelligent Systems. protégé. https://protege.stanford.edu/

[9] U. Yadav, G. S. Narula, N. Duhan, V. Jain, and B. K. Murthy, “Development and Visualization of Domain Specific Ontology using Protege,” *Indian Journal of Science and Technology*, vol. 9, no. 16, May 2016, doi: https://doi.org/10.17485/ijst/2016/v9i16/88524.

[10] M. Arenas and J. A. Pérez, “Querying semantic web data with SPARQL,” Symposium on Principles of Database Systems, Jun. 2011, doi: https://doi.org/10.1145/1989284.1989312.

[11] G. Reif, H. C. Gall, and M. Jazayeri, “WEESA,” Jan. 2005, doi: https://doi.org/10.1145/1060745.1060849.

[12] “Citywide crime statistics,” City Wide Crime Stats - NYPD, https://www.nyc.gov/site/nypd/stats/crime-statistics/citywide-crime-stats.page.

[13] P. D. (NYPD), “NYPD Shooting Incident Data (historic): NYC Open Data,” NYPD Shooting Incident Data (Historic) — NYC Open Data, https://data.cityofnewyork.us/Public-Safety/NYPD-Shooting-Incident-Data-Historic-/833y-fsy8.

[14] P. D. (NYPD), “NYPD hate crimes: NYC open data,” NYPD Hate Crimes — NYC Open Data, https://data.cityofnewyork.us/Public-Safety/NYPD-Hate-Crimes/bqiq-cu78.

[15] P. D. (NYPD), “NYPD Complaint Data Historic: NYC Open Data,” NYPD Complaint Data Historic — NYC Open Data, https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i.