Knowledge Graphs: Gaining Deeper Insight Into Open-Source Data

Exploring WMATA's Metrorail System

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Emmitt Sun James

November 5, 2024

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISORS

Rider Foley, Department of Engineering and Society

Briana Morrison, Department of Computer Science

Introduction

Over the past 15 years, the volume of open-source data has grown exponentially, driven by the widespread adoption of smartphones, social media, and open data initiatives (Statista & IDC, 2021). This surge in data creation is largely a result of global access to technology and the internet, with 67.5% of the population now online and 63.8% using social media (DataReportal et al., 2024). With individuals worldwide actively sharing information and generating content across a variety of platforms, data is produced at an unprecedented rate and scale. Social media platforms are particularly notable contributors to this exponential growth. Platforms such as Twitter and Facebook encourage users to share everything from daily life updates to opinions on global events, creating a highly dynamic and constantly updated information landscape. Additionally, many of these platforms focus not only on text-based posts but also on images and videos, further diversifying the data and creating expansive multimedia repositories.

This massive scale of data that is openly available brings significant opportunities. All of this data forms a diverse and complex dataset, providing real-time snapshots of human behaviors, attitudes, and events from around the world. Analysts and organizations across many different industries can use this data to gain a deeper understanding of various phenomena, from consumer behavior to political sentiment.

However, despite the accessibility and the potential of this data, there are many challenges that come with trying to utilize it. One of the biggest challenges in working with open-source data is handling the sheer volume of data that is available. With this massive volume of data, it is hard to filter, clean, and organize information effectively. Open-source data also tends to be largely unstructured, making it even harder to deal with (Krstic, 2024).

To address the complexity and challenges of open-source data, new approaches and tools are needed to effectively process and analyze this vast and varied information. Traditional methods of data analysis such as relational databases or flat-file processing struggle to capture the rich, interconnected nature of this data. This hinders researchers and analysts from obtaining a deeper understanding of interdependencies within the data, limiting its ability to help solve real-world problems. This technical project focuses on my internship this past summer where my team and I worked on a project centered around a data pipeline designed to ingest raw open-source data, process it, and put it into a knowledge graph.

Technical Topic

Traditional data storage and analysis methods, such as relational databases, are designed around rigid schemas and predefined relationships. While these systems excel at handling structured data with clear hierarchical relationships, they struggle with the dynamic and interconnected nature of open-source data (Kiff, 2024). Relational databases require predefined table structures and join operations that become increasingly complex and computationally expensive as the number of relationships grows. Similarly, flat-file processing systems, while simple to implement, lack the ability to efficiently represent and query complex relationships between data points.

Knowledge graphs offer a more flexible and powerful approach to handling complex data relationships. At their core, knowledge graphs are graph-based data structures where information is represented as nodes (entities) connected by edges (relationships). This structure allows for more natural representation of real-world relationships and enables more efficient querying of connected data. In our implementation, nodes represent various entities found in the open-source

data, while edges represent the relationships between these entities. An example of this would be the entities of “LeBron James” and “Bronny James” with a relationship of “Son of” in one direction and “Father of” in the other direction, as well as a relationship of “Teammate of” in both directions. Amazon Neptune, Amazon Web Services’ (AWS) fully managed graph database service, was used to create and manage our knowledge graph.

The project is composed of three main components: the data pipeline, the knowledge graph, and the web application, all of which integrate seamlessly to enable the collection and visualization of open-source data (see Figure 1). The data pipeline component handles the automated collection and preprocessing of open-source data from various sources. The process begins with data collection, where open-source data is periodically collected automatically. Then, the collected data is cleaned and standardized to ensure quality and consistency. Following this, entity and relationship recognition is performed using DBpedia Spotlight, an external natural language processing tool that identifies and annotates key entities—specifically people, places, and organizations. DBpedia Spotlight enriches these entities with additional data, including relevant properties and connections to other entities, which is crucial for building a comprehensive knowledge base.

In the knowledge graph component, the processed data is transformed to align with a predefined structure, known as an ontology, preparing it for insertion into our knowledge graph. This transformation ensures that the data follows a structured format compatible with our Amazon Neptune graph database. Once transformed, the data is seamlessly loaded into Neptune, where it powers the relationships and insights within the knowledge graph.

The web application includes both a frontend and a backend, creating a seamless user experience for interacting with the knowledge graph. The frontend allows users to explore and

navigate the data intuitively, while the backend acts as a “data shuttle”, efficiently retrieving and delivering relevant data from the knowledge graph to the user-facing interface. As users navigate to different parts of the application, the backend dynamically fetches the latest data from the knowledge graph, ensuring that users have access to up-to-date data.

The data pipeline was hosted entirely using various Amazon Web Services (AWS) services, a cloud computing platform that provides many different services for cloud-based data processing and storage options. Leveraging AWS allowed us to create a scalable and efficient pipeline, streamlining data ingestion and processing, while reducing the complexity of managing separate infrastructures

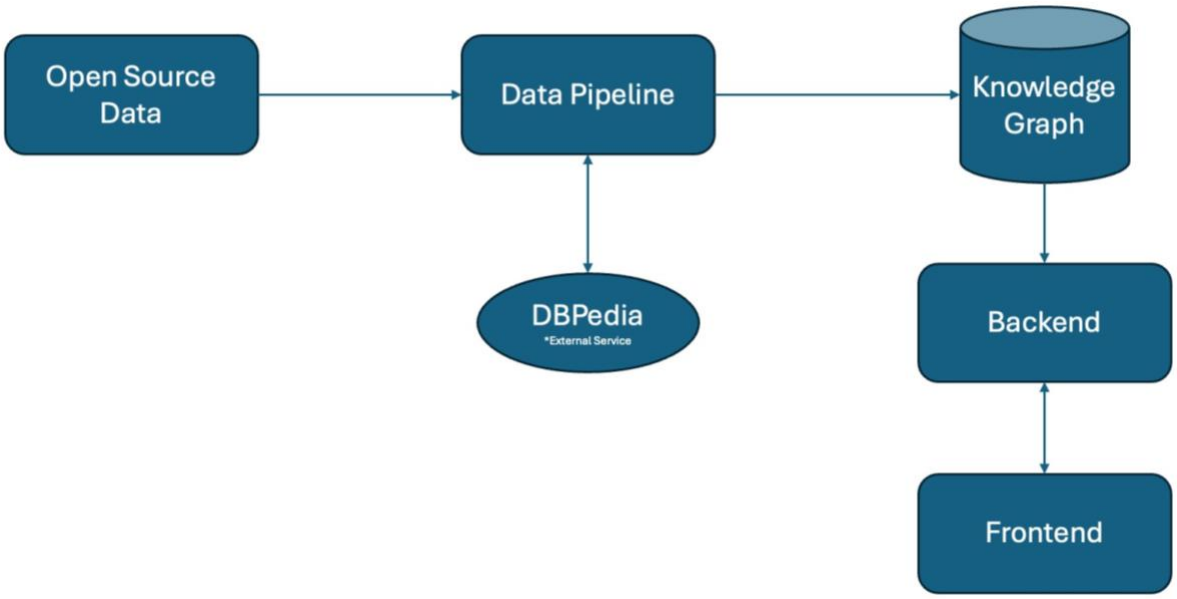

Figure 1. Project Flow

STS Topic:

While working on the technical project, my team and I had to be in the office 5 days a week. However, commuting to the office was not always an easy task, as heavy traffic and limited access to personal vehicles posed challenges for some team members. This led several of

us to rely on the public transit options provided by the Washington Metropolitan Area Transit Authority (WMATA), particularly the Metrorail system. The decision to rely on the Metro for our daily commutes highlights the critical role that public transportation infrastructure plays in shaping the lived experiences of those who interact with it. By applying Susan Star's infrastructure framework (1999), we can take a closer look at the complex socio-technical relationships that surround the Metrorail system.

One property of infrastructure that Star discusses is reach and scope, which emphasizes how infrastructure spans across a large spatial and temporal area (1999). Serving an average of 374,460 passengers per day across the District of Columbia and into the surrounding suburbs of Maryland and Virginia as well as being open most hours of the day (WMATA, 2024), Metrorail has become essential for connecting communities and facilitating the daily routines of commuters. Additionally, Metrorail connects to other transit options, extending its influence beyond just the DC area. This property of infrastructure has also shaped the mission of Metrorail, which has been steadily expanding its services to reach more and more communities throughout its history (WMATA, 2024).

Another property that Star discusses is that infrastructure becomes visible upon breakdown (1999). For regular riders, the Metrorail system typically functions as a background utility, taken for granted as a reliable form of transportation. However, when technical issues arise — be it delays, breakdowns, or track maintenance — the system's presence suddenly becomes very visible. Breakdowns often lead to frustration and a reevaluation of its reliability, underscoring how the invisibility of well-functioning infrastructure is a mark of its success (Mo et al., 2022). This idea shapes a major goal for WMATA: to provide consistent, reliable transportation with minimal disruptions (WMATA, 2024). Given that breakdowns make the

system's flaws more apparent, WMATA must prioritize reliability, or risk commuters turning to alternative options. Whenever large breakdowns do happen, such as the recall of all 7000-series railcars back in 2021, it causes major service issues which are very visible (Duncan, 2021).

Star also discusses how infrastructure typically follows an embodiment of standards, allowing for interactions between people and technology to play out in a predictable way (1999). This principle is shown in WMATA's Metrorail service standards, which highlight all aspects of the system ranging from train frequencies to equity and non-discrimination policies (WMATA, 2021). Furthermore, the system must adhere to safety standards which are enforced by the Washington Metrorail Safety Commission, encouraging WMATA to invest in safety features throughout the system. On top of this, the infrastructure of the system itself follows many standards such as track gauge, allowing for all models of Metrorail train cars to run on the same tracks with a consistent user experience (Keffler, 1985). These standards shape not only the technical operation of the system but also the social interactions that occur within it, establishing predictable norms and expectations for passengers.

Research Question

WMATA's Metrorail system provides valuable benefits to both individual riders and the broader community, including reduced emissions and enhanced socioeconomic mobility (WMATA, 2024, 2023). However, high ridership levels are needed not only improve the system's financial viability but also enable WMATA to further invest in service enhancements and community initiatives. Given these considerations, a key question emerges: How can WMATA better utilize its funding to increase ridership on the Metro?

The primary approaches I will take to answering this question will be interviews with professionals and analysis of existing WMATA studies. First, structured interviews will be conducted with WMATA employees across various roles, including station managers and administrators. These interviews will provide valuable insider perspectives on operational challenges and resource allocation. Second, I will analyze existing WMATA studies and reports to gain insights into certain domains like current initiatives and metrics like ridership trends. These research methods will provide a comprehensive understanding of WMATA's operations, challenges, and opportunities.

The data collected through these methods will be analyzed using Susan Star's properties of infrastructure to develop well-informed recommendations for how WMATA can better utilize its funding to attract and retain riders. This holistic approach will provide a comprehensive understanding of the challenges and opportunities in increasing Metrorail ridership.

Conclusion

The growing volume of open-source data presents both opportunities and challenges for meaningful analysis and insight generation. For the long-term sustainability of open-source data analysis, the knowledge graph-based data pipeline developed in this technical project serves as a promising. By providing a robust alternative to traditional data analysis methods, this solution has shown that knowledge graphs can be applied to handle open-source data in an efficient and scalable manner, enabling better data organization and insight extraction.

The STS deliverables of this research provide valuable insights into how Metrorail interacts with society, through the use of Susan Star's infrastructure framework (1999). Research will be conducted and analyzed using Star's infrastructure framework to understand how

WMATA can better allocate its resources to improve the Metrorail system. The findings from interviews with professionals and the analysis of WMATA studies will provide actionable insights for WMATA to enhance ridership and system effectiveness. The conclusions may help to drive increased ridership, and therefore an increased effectiveness of WMATA's Metrorail system.

Resources

DataReportal, Meltwater, & We Are Social. (2024, October 23). *Number of internet and social media users worldwide as of October 2024 (in billions) [Graph]*. Statista. https://www.statista.com/statistics/617136/digital-population-worldwide/

Duarte, F. (2023, March 16). Amount of Data Created Daily (2024). *Exploding Topics*. https://explodingtopics.com/blog/data-generated-per-day

Duncan, I. (2021, December 29). Metro safety commission orders cars out of service, saying agency didn't follow terms of plan. *Washington Post*. https://www.washingtonpost.com/transportation/2021/12/29/metro-7000-series-removed-service/

Keffler, A. (1985). *The Evolution of Washington Metro's Track Standards*. https://onlinepubs.trb.org/Onlinepubs/trr/1985/1006/1006-006.pdf

Kiff, L. (2024, February 16). Exploring the Revolution: Graph Databases in Modern Data Management. *Tom Sawyer Software*. https://blog.tomsawyer.com/knowledge-graph-vs-graph-databases

Krstic, B. (2024, June 19). Social Media: Unstructured Data and How to Utilize It. *Jatheon*. https://jatheon.com/blog/is-social-media-unstructured-data/

Mo, B., Von Franque, M. Y., Koutsopoulos, H. N., Attanucci, J. P., & Zhao, J. (2022). Impact of Unplanned Long-Term Service Disruptions on Urban Public Transit Systems. *IEEE Open Journal of Intelligent Transportation Systems*, *3*, 551–569. https://doi.org/10.1109/OJITS.2022.3199108

Petkova, G. (2024, January 26). Knowledge Graphs: Redefining Data Management for the Modern Enterprise. *Ontotext*. https://www.ontotext.com/blog/knowledge-graphs-redefining-data-management-for-the-modern-enterprise/

Star, S. L. (1999). The Ethnography of Infrastructure. *American Behavioral Scientist*, *43*(3), 377–391. https://doi.org/10.1177/00027649921955326

Statista & IDC. (2021, June 7). *Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025 (in zettabytes) [Graph]*. Statista. https://www.statista.com/statistics/871513/worldwide-data-created/

Stegeman, J. (2024, July 22). *What is a knowledge graph?* Neo4j. https://neo4j.com/blog/what-is-knowledge-graph/

WMATA. (2021). *Metrorail Service Standards*. WMATA. https://www.wmata.com/initiatives/plans/upload/Metrorail-Service-Standards.pdf

WMATA. (2023a). *Your Metro, The Way Forward Strategic Transformation Plan*. WMATA. https://wmata.com/initiatives/strategic-plan/upload/230314_STP_Report.pdf

WMATA. (2023b, May 19). *Stations*. WMATA. https://wmata.com/rider-guide/stations/index.cfm

WMATA. (2024a). *2024 Benefits of Transit Report*. WMATA. https://wmata.com/about/news/upload/BenefitsOfTransit_Report_240304_Web.pdf

WMATA. (2024b). *FY2025 BUDGET*. WMATA. https://wmata.com/initiatives/budget/upload/Remediated-FY2025-Approved-Budget-FINAL.pdf

WMATA. (2024c, September 29). *Metrorail*. WMATA. https://wmata.com/service/rail/index.cfm

WMATA. (2024d, November 1). *Metrorail Ridership Summary*. WMATA. https://wmata.com/initiatives/ridership-portal/Metrorail-Ridership-Summary.cfm

Wu, S., Zhuang, Y., Chen, J., Wang, W., Bai, Y., & Lo, S. (2019). Rethinking bus-to-metro accessibility in new town development: Case studies in Shanghai. *Cities*, *94*, 211–224. https://doi.org/10.1016/j.cities.2019.06.010

Xu, X., Lu, Y., Wang, Y., Li, J., & Zhang, H. (2020). Improving Service Quality of Metro Systems—A Case Study in the Beijing Metro. *IEEE Access*, *8*, 12573–12591. https://doi.org/10.1109/ACCESS.2020.2965990