PROCEEDINGS OF SCIENCE

Leveraging Knowledge Graph-Enhanced RAG and LLMs for Historical Archival Analysis: A Case Study of State of Maryland's Dataset Collections

Rajesh Kumar Gnanasekaran,[a,*] **Richard Marciano**[b] **and Chris Haley**[c]

[a] *University of Maryland,*
College Park, MD, USA
[b] *University of Maryland,*
College Park, MD, USA
[c] *Maryland State Archives,*
Annapolis, MD, USA

E-mail: rgnanase@umd.edu, marciano@umd.edu, chris.haley@maryland.gov

---

[*] Speaker

© Copyright owned by the author(s) under the terms of the Creative Commons
Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0).

Integrating Artificial Intelligence in digital humanities has created unprecedented opportunities for analyzing historical archives. Building upon established work with Maryland State Archives (MSA)' Legacy of Slavery (LoS) collections, this research proposes an innovative approach combining Knowledge Graph-based Retrieval Augmented Generation (KG-RAG) with Large Language Models (LLMs) to analyze three dataset collections: Certificates of Freedom, Domestic Traffic Advertisements, and Manumissions. These collections are historically rich and focused on uncovering the narratives of individuals who resisted enslavement in Maryland, USA. This project introduces a novel architecture that enhances traditional Generative AI RAG systems by incorporating prompt-engineered reasoning over a knowledge graph instead of relying on vector-based semantic similarity. Unlike conventional RAG approaches that embed user queries and documents into a shared vector space for retrieval, this system uses structured Cypher queries generated via prompt templates to interact directly with a Neo4j-based knowledge graph. This design allows for precise symbolic reasoning over richly interconnected historical data, enabling nuanced natural language exploration without the need for approximate embedding-based matching. The system employs a three-layer architecture: a knowledge graph layer mapping relationships between entities across collections using Neo4j, an RAG layer augmented through prompt-driven Cypher generation and contextual retrieval, and an LLM layer for natural language synthesis based on grounded graph responses. This study builds upon earlier iterations of ChatLoS—a simple RAG chatbot and an agentic CSV-based version—by structurally transforming the retrieval method to support cross-collection linkage and entity-aware responses. Rather than stacking redundant LLM layers, each iteration addresses specific limitations with simple RAG, and csv-agent AI. By eliminating the need for specialized database knowledge or understanding of archival organization systems, the interface significantly improves accessibility which is the main goal of the LoS project. Additionally, this prompt-engineered KG-RAG architecture advances AI-enabled scientific workflows by leveraging specialized prompt engineering patterns for cross-collection analysis and by preserving the interpretability and provenance of historical evidence. It enhances the trustworthiness and accuracy of insights by grounding responses in verified relationships rather than probabilistic approximations. While the system may at times surface structured results such as counts or connections, these outputs are semantically rich indicators of deeper historical narratives, enabling natural language interactions that democratize access to complex archives. To assess the utility and usability of these tools from a domain expert's perspective, a qualitative user study was conducted with the Director of the LoS project. The study revealed key themes: current tools at MSA are rigid and siloed, requiring users to have detailed schema knowledge; in contrast, ChatLoS significantly lowers access barriers by supporting natural language queries and conversational refinement. The KG-RAG version was particularly praised for its ability to trace individuals across legal and commercial records and to enhance trust via explainable connections and citations. This study concludes by listing down the limitations and proposing future participatory evaluations with descendant communities and public-facing users, to ensure the design and deployment of AI tools for archival research are culturally responsive, ethically grounded, and historically contextualized.

International Symposium on Grids and Clouds (ISGC2025)
16 -21 March, 2025
Academia Sinica Computing Centre (ASGC), Institute of Physics, Academia Sinica Taipei, Taiwan

1. Introduction

Cultural archives serve as custodians of collective memory, preserving narratives that shape historical understanding and contemporary identity [1]. Traditional archival practices rooted in manual cataloging and reading struggle to keep pace with the scale and complexity of digital-era collections. Many archives have become essentially large troves of data preserved but not readily accessible to users [2]. For example, digitized collections like the State of Maryland's Legacy of Slavery (LoS) project contain extensive data on enslaved individuals and emancipated communities. However, unlocking insights from such data requires computational analysis beyond conventional methods. This gap between archival wealth and accessibility has been widely recognized as a critical challenge [2] [3]. At the same time, the rise of data science and artificial intelligence (AI) offers promising avenues to bridge this gap. In particular, computational methods can dramatically enhance how we process, analyze, and engage with archival materials [3].

AI has proven "*essential for cleaning, exploring, and visualizing archival and special collections*" [3]. These technologies can facilitate what was once impossible: conversing with an archive in natural language. However, realizing this potential in archives comes with its challenges. AI systems (especially generative AI (Gen-AI) [4] powered solutions like large language models (LLMs) [5]) carry risks of ethical pitfalls – for example, they might introduce inaccuracies or biases ("hallucinations" [6]) when applied to sensitive historical data [7]. The convergence of these trends has created an urgent need for research that develops AI-driven tools ethically and responsibly to enhance archival access by overcoming these challenges. Historical archives, such as the LoS collections, contain rich but fragmented data spread across multiple record sets. Unlocking cross-collection insights (e.g., tracing an enslaved individual from a sale advertisement to their manumission and freedom) is challenging with traditional archival databases. Users (researchers, genealogists, archivists) must manually search separate datasets and piece together narratives, a process that is time-consuming and requires expert knowledge of the data schema.

Recent advances in Retrieval-Augmented Generation (RAG) with LLMs offer a promising avenue for more accessible archival analysis by allowing users to ask questions in natural language and receive synthesized answers grounded in relevant records. However, standard RAG systems typically rely on vector similarity search over documents and struggle to capture complex relationships or perform reasoning that spans multiple collections [8]. In response, we propose a novel approach that augments RAG with a Knowledge Graph (KG) to better model the people, places, and events in the LoS archives and improve the accuracy and transparency of the AI's responses.

In earlier work with ISGC [9], we developed ChatLoS, a prototype chatbot using RAG (with OpenAI's GPT-3.5) on a single LoS dataset (Domestic Traffic Ads). It demonstrated the feasibility of conversational queries over archival data, but also highlighted limitations: it could only retrieve from one collection at a time with a small context window (couldn't fit a full dataset in its context window), could not perform aggregate queries (e.g., aggregate counts or statistics), and lacked ability to traverse relationships between datasets (e.g., linking an advertisement to a related manumission record). These issues mirror broader challenges noted in literature i.e., baseline RAG struggles to "connect the dots" between disparate pieces of information and to synthesize holistic answers that require traversing linked data [8].

KGs have emerged as a potential solution to provide LLMs with structured context and reasoning pathways beyond simple text matches. For example, [8]'s approach showed that incorporating an LLM-generated KG can substantially improve question-answering on private datasets, enabling the LLM to ground its answers in the graph and even provide provenance links to sources [8]. Similarly, in the cultural heritage domain, KGs have been used to interlink records and reveal hidden connections [10]. These successes suggest that a knowledge graph-enhanced RAG (KG-RAG) could address our use case's needs: by unifying multiple LoS datasets in a graph, we can help the LLM traverse cross-collection links (e.g. person-to-record relationships) and handle quantitative or constraint-based queries via cypher graph queries, while grounding answers in a network of verifiable facts.

In this paper, we present a case study of a prompt-engineered KG–Enhanced RAG system for a three sample set of the LoS dataset collections. We design and implement a KG-RAG architecture that combines Neo4j KGs with an LLM-based chatbot to enable natural language queries across three key LoS collections (Certificates of Freedom (CoF), Domestic Traffic Ads (DTAs), and Manumissions). The system dynamically translates user questions into graph queries (Cypher) based on prompt-engineered templates to retrieve multi-hop context and uses an LLM to generate a narrative answer. We evaluate and illustrate the system's capabilities with example queries, demonstrating how the KG enhancement enables cross-collection tracing of individuals, aggregate analysis, and more explainable answers. We also capture and produce a qualitative, single-participant case study with the Director of MSA's LoS project (Chris Haley) to gather preliminary expert feedback on the system by comparing the archival access methods using traditional data visualization tool (Tableau in this case), simple RAG ChatLoS, csv-agentic ChatLoS and the KG-RAG enhanced ChatLoS. This user study provides insights into the perceived utility, usability, and trustworthiness of the KG-RAG approach from an archival practitioner's perspective. We also discuss how our findings align with emerging research in using KGs for explainable and domain-specific AI [10], and outline a roadmap for a more formal, community-based evaluation in future work.

Overall, this work lies at the intersection of digital humanities, cultural heritage informatics, and AI. It advances the state-of-the-art by showing how KGs can be tightly integrated with LLMs to provide richer, more trustworthy conversational access to archival collections. This approach aims to democratize access to complex historical archives – allowing users to ask nuanced questions and discover linked historical narratives without requiring specialized query skills – while ensuring the AI's answers remain grounded in the archival records' context and provenance. To address this project's specific objectives mentioned above, we formulated the following research question.

- RQ1: How does the integration of Knowledge Graph-based Retrieval Augmented Generation (KG-RAG) with LLMs enhance the analysis and accessibility of interconnected historical archives, specifically the LoS collections at the MSA?

2. Background & Significance

Since its initiation in the fall of 2001, the MSA's project, officially named the Study of the LoS in 2005, has focused on uncovering the narratives of individuals who resisted enslavement in Maryland, USA. Utilizing various sources like court records, laws, newspapers, and maps, the MSA staff aimed to highlight the unrecognized "heroes" of slave resistance and create comprehensive case studies. This endeavor has culminated in a robust online database containing over 400,000 records, including DTA, manumissions, CoF records, and federal census data, providing a detailed view of anti-slavery movements across fourteen of the state's antebellum counties. Several prior works have been performed on these dataset collections for various purposes, as detailed in [11], [12], [13], [14]. For this project, we will be using the three datasets - CoF, Manumissions and the DTA datasets.

2.1 Domestic Traffic Ads Dataset

The DTA dataset comprises records that contain details on the domestic traffic advertisements placed in public newspapers for the interstate and intrastate trade of enslaved men, women, women, and children. This dataset comprises digitally transcribed data of advertisements for enslaved individuals placed in Maryland newspapers over a 40-year period, from March 3, 1824, to April 30, 1864. It includes crucial metadata fields such as advertisement date, enslaved person's name, slave owner's name, source publishing this ad, county, location, enslaved age, gender, number of people being sold, terms of sale, and specified skills. Originally, about 2100+ advertisement records were found in MSA's digital database. After performing preliminary data exploration and a cleaning process, for the purpose of this exploratory case study, we created a test dataset based on a subset of the records covering the first 10 years of the collection (from 1824 to 1834), which resulted in over a third (35%) of the dataset or 764 records.

2.2 Manumissions

A manumission record is the legal document issued often by the slave owners freeing a slave in the presence of a witness. These manumission records are one of the records that the enslaved people produce at the counties to get a certificate of freedom document. The digitally transcribed manumissions dataset has about 7200+ records.

2.3 Certificates of Freedom Dataset

A CoF is a legal document issued by the State of Maryland from their counties through the clerks to an enslaved person indicating that they have been freed and they have to carry this document with them as an identity or proof of freedom. The CoFs were handwritten documents containing general biographic, demographic, and descriptive information about the enslaved person. There are about 23000 records digitally transcribed as part of this dataset and is used as-is in this study.

3. Related Work

A knowledge graph (KG) is a structured representation of facts where entities (nodes) are connected by relationships (edges), often following ontologies or schemas. KGs have been used in cultural heritage contexts to interlink data across collections and provide rich context for items. For example, the Heritage Connector project built a KG linking museum collection records with external sources like Wikidata, which "*allowed visualization and exploration of collections in entirely new ways*" by highlighting connections between objects, people, and concepts that were not obvious before [3]. Such interlinking transforms isolated records into a network of knowledge, enabling users to traverse from one entity to related entities (e.g., an artist to their artworks across museums, or from a historical person to all archival records mentioning them). Another critical aspect of the accuracy and correctness of the responses by Gen-AI tools using unique retrieval mechanisms purely depends upon the domain knowledge of information that's shared with the LLMs. Many researchers and scientists believe that the data supplied to these LLMs should be prepared and pre-processed to be made *"AI-Ready"* as indicated by [15].

Additionally, in digital archives, KGs can address the limitation of "*item-centric*" views. Traditional archive databases show search results as lists of items with metadata [16]. In contrast, a KG-driven interface can show how items relate (chronologically, geographically, by topic, etc.), supporting users in browsing related items and synthesizing narratives [16]. Khoo et al. (2024) demonstrated this by developing a visualization interface for archives based on a KG; users could click on a person in one record and immediately see other records connected to that person, or follow links to events and places related to them [16]. The user study indicated that this helped in forming a more holistic understanding of the content, as users were not confined to one record at a time [16].

For this research, the interest in KGs is two-fold. First, as a tool for enriching archival data: by extracting entities (people, places, dates) from archival records and connecting them, we create a semantic layer that AI can use. Several published studies have shown the use of KGs in their problems in combination with LLMs to successfully get better results than with just the use of LLMs and the embeddings based RAG techniques [17] [18] [19] [20] [21]. Second, KGs are central to addressing the context deficit of LLMs. An LLM generates answers from patterns in text, but it doesn't "*know*" a fact in a robust way that it can cite or justify. If we integrate an LLM with a KG, it can retrieve relevant nodes or subgraphs to ground its responses. The literature on retrieval-augmented LLMs suggests they can significantly reduce hallucination and improve factual accuracy by anchoring generation on retrieved information [22].

Our work contributes to this discourse by evaluating an archivist's perspective on whether showing linked records and data paths in an answer (e.g. enumerating the steps in an enslaved person's journey across documents) improves credibility. In summary, our system builds upon these lines of research by combining a KG of historical slave records with an LLM in a RAG framework. To our knowledge, this is one of the first applications of KG-RAG in this context. Next, we describe the design of our system and the methodology of our evaluation.

4. Research Methodology & Questions

Our research methodology uses the design science approach in the development of the KG-RAG enhanced ChatLoS and an exploratory qualitative user evaluation through a case study interview. To address the RQ1, we designed the system architecture and KG construction as below, then discuss the results, limitations and future work.

4.1 System Architecture and KG Construction

The proposed system comprises three coordinated layers: a KG layer, a RAG layer, and a LLM Question-Answering layer. This architecture is designed not as a traditional embedding-based RAG layer, but as a prompt-engineered pipeline that delegates symbolic reasoning to a Neo4j-based KG, thereby enhancing trust, traceability, and contextual depth without introducing human-opinionated bias.

Knowledge Graph Layer: The KG is implemented in Neo4j as a property graph and serves as the structured semantic backbone for exploring the three LoS datasets. We integrated the three LoS datasets: (1) *DTA*, newspaper listings offering enslaved individuals for sale, (2) *Manumission Records*, legal records of emancipation, and (3) *CoF*, formal documents certifying freedom, often following manumission into this KG. Each record type maps to a corresponding node label (e.g., `DomesticTrafficAd`, `ManumissionRecord`, `CertificateOfFreedom`), with entity-level nodes created for shared actors such as `EnslavedPerson` and `SlaveOwner`. We modeled historically grounded relationships between these nodes (e.g., `LISTED_FOR_DOMESTIC_TRAFFIC_SALE_IN`, `MANUMITTED_THROUGH`, `WAS_GRANTED_FREEDOM_THROUGH`) to trace life events across datasets. All linkages are constructed from data-backed properties (e.g., names, dates, locations), and no inferred connections or manual opinion-based mappings were introduced, ensuring that the KG remains a faithful, verifiable source-of-truth grounded solely in archival records. We chose Neo4j for its capability to store property graphs and execute Cypher queries, and we populated it using a script-based import for nodes/edges (e.g., loading CSVs of each dataset and then linking records by shared attributes where available such as enslaved names, slave owner names, county names, and gender). Tables 1 and 2 show the list of nodes, their properties and their relationships between them created for the KG-RAG enhanced ChatLoS. Figure. 1 shows the nodes and relationships schema on neo4j built for ChatLoS.

Graph-RAG Layer (Cypher-based Retrieval): Unlike standard RAG systems that retrieve semantically similar passages via embedding-based vector search, this system operates on a Cypher-queryable KG that enables structured reasoning. When a user submits a natural language question, the system prompts the OpenAI's GPT-4o model to generate a Cypher query tailored to the schema (e.g., counting ads by county/date, or tracing an individual's pathway from sale to freedom). For example, if the user asks *"How many ads were placed in Dorchester County in the 1820s?"*, the LLM might generate a Cypher query that counts `DomesticTrafficAd` nodes filtered by county=Dorchester and date range 1820–1829. A curated prompt template guides the LLM's Cypher generation to ensure syntactic and semantic alignment with the KG. The system executes the Cypher on Neo4j, retrieves the results (which could be a set of records or aggregate values), and then another prompt is used to have the LLM compose a final answer in natural language, using the query results as its grounding. This design is inspired by recent work using LLMs to translate natural language to structured queries for graphs [20]. It allows complex multi-hop questions to be answered by first pulling the relevant subgraph or data from Neo4j, instead of relying purely on vector similarity. This design is intentionally non-layered. The KG-RAG does not stack multiple vector RAGs or agentic tools, but instead replaces embedding retrieval with structured querying, avoiding spurious matches and improving explainability. In edge cases where Cypher fails, an error message indicating that there were no matching records or a fallback text message is displayed.

LLM QA Layer: The final response layer employs GPT-4o to synthesize a human-readable answer from the retrieved graph results. Prompts are engineered to encourage structured, citationrich output (e.g., narrating an individual's freedom journey with date-stamped events and linked archival sources). Because the LLM operates on a verified Cypher output and is never exposed to raw training embeddings from the corpus, the generation remains grounded and interpretable. In summary, the KG provides a "source of truth" graph that the LLM can query and reason over, thereby constraining the generation to verified archival data and complex relationships within. This KG-RAG approach differs from traditional RAG in both design and philosophy. It uses a schema-first, symbolic method that reflects the structure of historical records and reduces dependency on opaque embedding spaces. It does not introduce authorial bias or speculative inferences. Instead, it empowers users—especially those unfamiliar with database query languages—to explore complex, cross-dataset historical narratives through a natural language interface backed by transparent graph-based reasoning.

Table 1: Node Types and Properties in the ChatLoS Neo4j KG

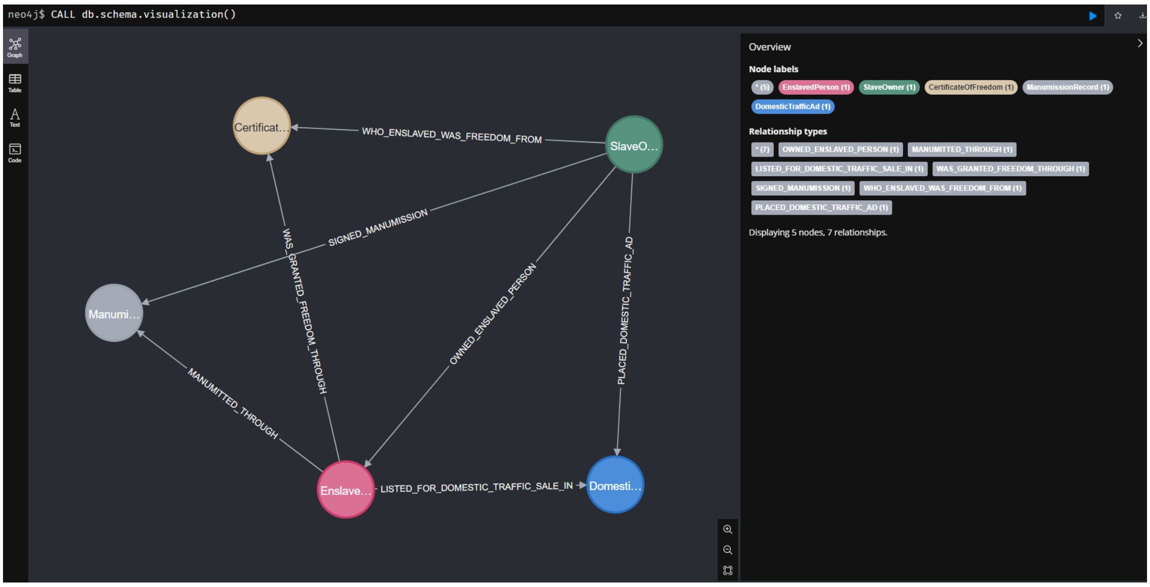
Node Label	**Representative Properties**
`EnslavedPerson`	first_name, last_name, sex, age, height, complexion, prior_status, alias
`SlaveOwner`	full_name, owner_first_name, owner_last_name, owner_middle_name
`ManumissionRecord`	DateManumitted, DateRecorded, Source, Witness_FullName, county_name
`CertificateOfFreedom`	date_freed, series, item, folder, document, page, Freed_FirstName, Freed_LastName
`DomesticTrafficAd`	Sale_Date, Newspaper_code, Trade_reason, Ad_Image_Metadata, skills_specified

4.2 User Study Design

We conducted a preliminary user evaluation in the form of a qualitative case study with a single expert participant: Chris Haley, Director of the Maryland State Archives' Legacy of Slavery program. The choice of a single-participant case study was deliberate for this formative stage – our goal was to gather in-depth feedback from a domain expert (and key stakeholder) to guide further development, rather than to generalize findings. This aligns with a formative evaluation approach, treating the study as an opportunity to uncover issues and opportunities through rich qualitative insight.

Table 2: Relationship Types in the ChatLoS Neo4j KG

Relationship Type	**Source Node**	**Target Node**
`OWNED_ENSLAVED_PERSON`	SlaveOwner	EnslavedPerson
`PLACED_DOMESTIC_TRAFFIC_AD`	SlaveOwner	DomesticTrafficAd
`LISTED_FOR_DOMESTIC_TRAFFIC_SALE_IN`	EnslavedPerson	DomesticTrafficAd
`SIGNED_MANUMISSION`	SlaveOwner	ManumissionRecord
`MANUMITTED_THROUGH`	EnslavedPerson	ManumissionRecord
`WAS_GRANTED_FREEDOM_THROUGH`	EnslavedPerson	CertificateOfFreedom
`WHO_ENSLAVED_WAS_FREEDOM_FROM`	CertificateOfFreedom	SlaveOwner



```
neo4j$ CALL db.schema.visualization()
```

Displaying 5 nodes, 7 relationships.

Figure 1: Entity-Relationship Schema of the ChatLoS KG in Neo4j

Procedure: The session was structured as a semi-structured interview (1 hour) with questions organized into sections corresponding to the evolution of the ChatLoS tool. We first discussed the Current State of LoS data access, to understand baseline workflows and pain points. Next, we introduced ChatLoS v1 (simple RAG-based chatbot) [9] and asked for impressions – e.g., *"Does the conversational format feel approachable?," "How helpful is it that no knowledge of schema is required?"* Then we demonstrated ChatLoS v2 (CSV-Agentic AI for data aggregation analysis) by comparing it with a visualization tool, Tableau and an equivalent natural language query for aggregate questions (like counting ads by county or date) [9]. Finally, we presented ChatLoS v3 (KG-RAG enhanced) – our primary contribution – and walked through examples of cross-dataset questions it can answer. We encouraged open feedback throughout, with prompts about explainability (e.g., *"Does seeing the paths between entities add credibility or trust?"*) and future use cases (e.g., *"What questions would you want to explore across datasets that current systems can't support?"*). The interview was conducted via teleconference, recorded, and transcribed for analysis with participant consent.

Data Analysis: We analyzed the interview transcript using manual thematic coding. Two researchers (including the lead author) reviewed the transcript and independently noted recurring themes or salient points in the participant's responses. We then discussed and consolidated these into a set of key themes, such as Current System Limitations, Perceived Benefits of Chatbot Interface, Concerns about Accuracy/Consistency, Value of KG Context, and Future Opportunities. Given the single-participant nature, we did not calculate inter-rater reliability; instead, we collaboratively interpreted the data, which is appropriate for an exploratory case study aiming for depth over breadth. Where relevant, we extracted illustrative quotes from the participant to include in our results (with light editing for clarity). The findings from this interview are intended to inform design and will serve as a foundation for more extensive evaluations. In future work, we plan to develop a more formal user study (with multiple participants and possibly a task-based evaluation), guided by this initial feedback – we explicitly separate that future evaluation plan from the present study to acknowledge that the current findings are preliminary.

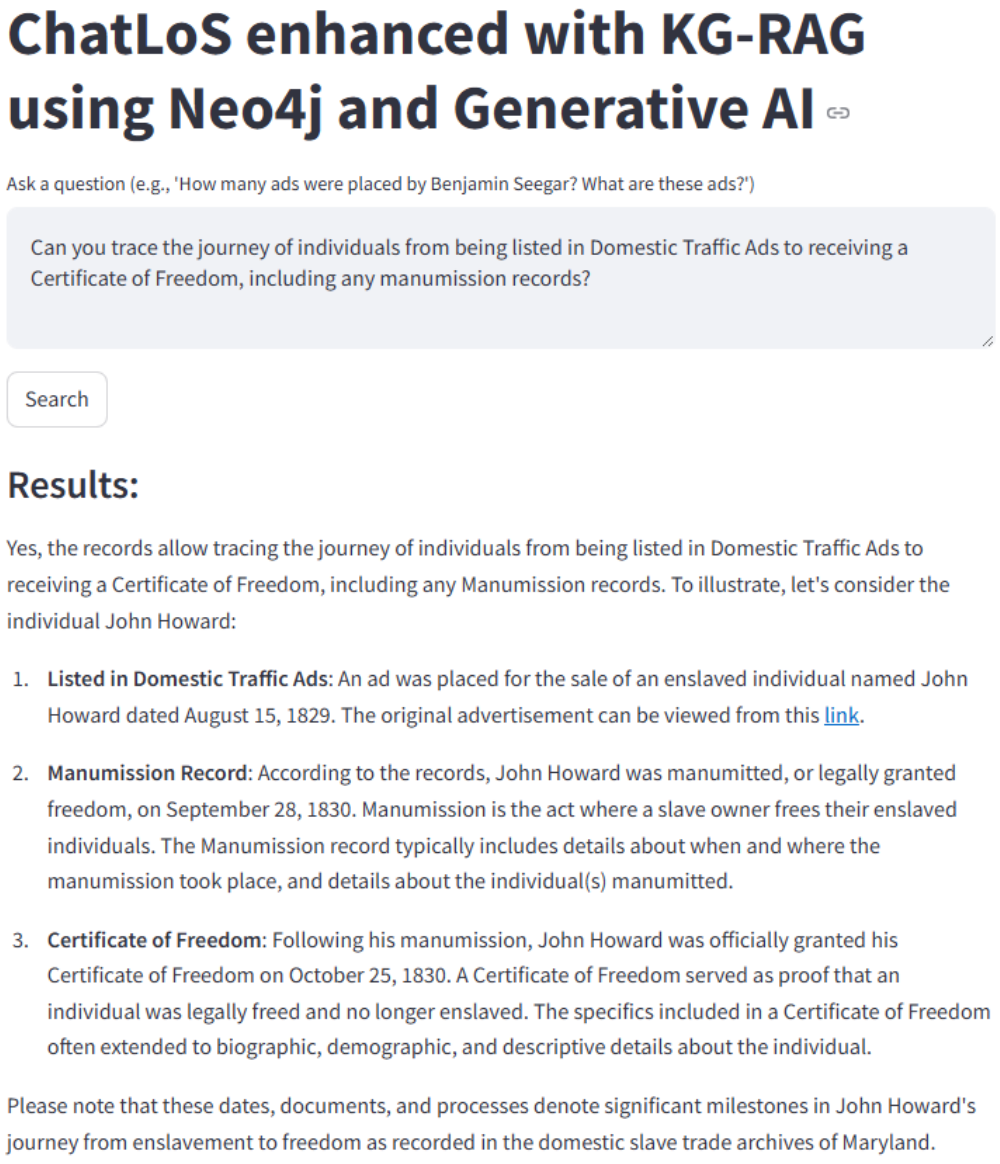
5. Results

We present our results in two parts: (1) System Performance & Examples, illustrating how the KG-enhanced ChatLoS addresses key research questions (cross-collection queries, aggregate reasoning, explainability) with example interactions; and (2) User Feedback from the case study, summarizing the expert's insights on the system's utility, limitations, and potential.

5.1 System Behavior with KG-RAG Enhancement

To evaluate the system's capabilities, we posed a series of questions representative of typical and challenging user inquiries, and we examined the answers generated by ChatLoS with the KG-RAG pipeline. We highlight two illustrative examples here (with screenshots).

Cross-Collection Tracing Query: One powerful use case enabled by the KG is tracing an individual's journey through multiple record types. We asked the system: *"Can you trace the journey of individuals from being listed in Domestic Traffic Ads to receiving a Certificate of Freedom, including any manumission records?"* – essentially querying if the system can follow an enslaved person's path from sale to freedom. The chatbot's response as shown in 2 demonstrates an ability to perform this multi-hop reasoning. The system chose an example individual (John Howard) and outlined: (1) Listed in a *DTA* – an ad dated August 15, 1829 for John Howard (with a hyperlink to the original advertisement record as shown in Figure. 3); (2) *Manumission Record* – noting John Howard was manumitted on Sept 28, 1830 (with contextual explanation of manumission); (3) *CoF* – stating he was granted a CoF on Oct 25, 1830. The answer explicitly references the linked records (e.g., "the original advertisement can be viewed from this link") and provides a narrative connecting them. This showcases the system's ability to traverse the KG (ad → manumission → freedom) and present a coherent story that spans collections. The above result illustrates how the KG provides the contextual backbone for answering what would be a very complex query for a standard RAG system. Without the graph, an LLM would have to somehow know that *"John Howard"* in one dataset is the same person in another – a task prone to error. Here, the graph query was able to find a matching individual node connected to records of all three types, and the LLM then described



ChatLoS enhanced with KG-RAG using Neo4j and Generative AI

Ask a question (e.g., 'How many ads were placed by Benjamin Seegar? What are these ads?')

Can you trace the journey of individuals from being listed in Domestic Traffic Ads to receiving a Certificate of Freedom, including any manumission records?

Results:

Yes, the records allow tracing the journey of individuals from being listed in Domestic Traffic Ads to receiving a Certificate of Freedom, including any Manumission records. To illustrate, let's consider the individual John Howard:

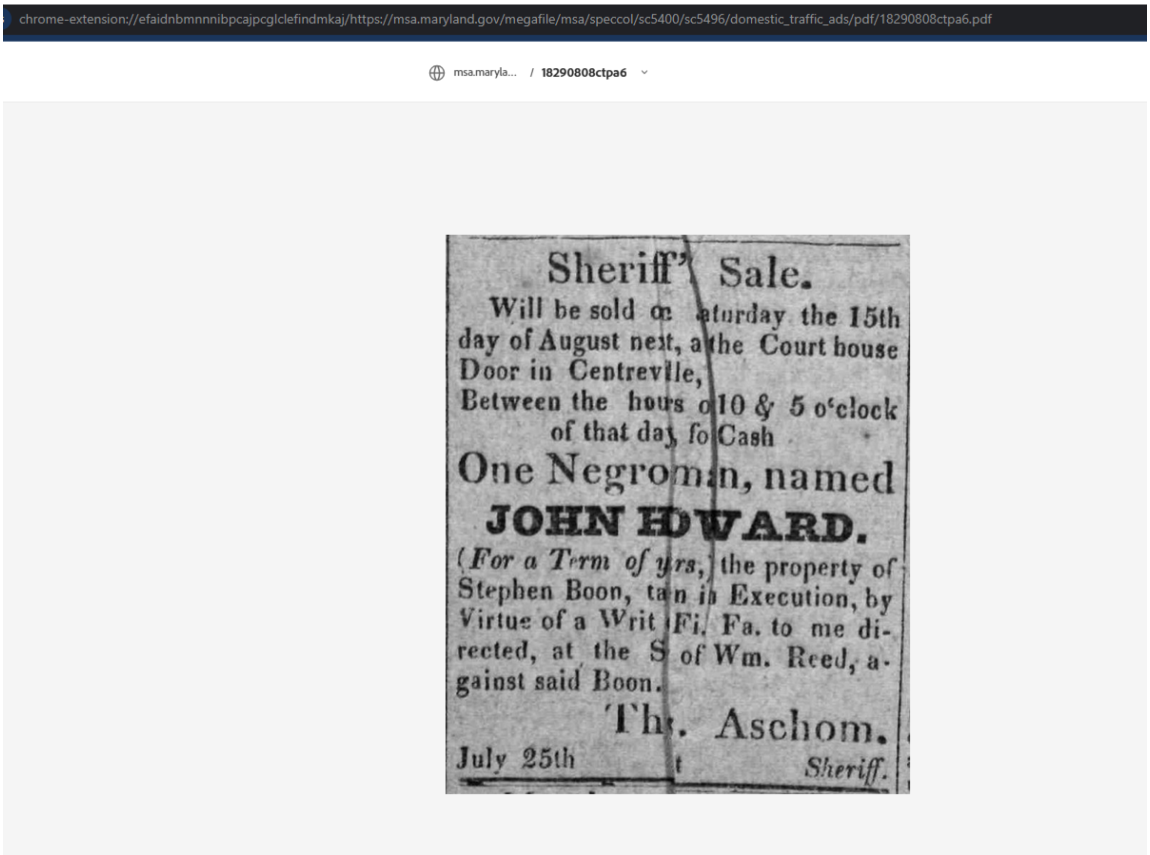
1. **Listed in Domestic Traffic Ads**: An ad was placed for the sale of an enslaved individual named John Howard dated August 15, 1829. The original advertisement can be viewed from this link.
2. **Manumission Record**: According to the records, John Howard was manumitted, or legally granted freedom, on September 28, 1830. Manumission is the act where a slave owner frees their enslaved individuals. The Manumission record typically includes details about when and where the manumission took place, and details about the individual(s) manumitted.
3. **Certificate of Freedom**: Following his manumission, John Howard was officially granted his Certificate of Freedom on October 25, 1830. A Certificate of Freedom served as proof that an individual was legally freed and no longer enslaved. The specifics included in a Certificate of Freedom often extended to biographic, demographic, and descriptive details about the individual.

Please note that these dates, documents, and processes denote significant milestones in John Howard's journey from enslavement to freedom as recorded in the domestic slave trade archives of Maryland.

Figure 2: KG-RAG ChatLoS Tracing an individual's path to freedom

those connections. Notably, the answer also took steps toward explainability by citing the existence of an original advertisement link. This aligns with the MSA director's desire to see sources and relationships in answers to build trust. It also provides an example of how the system mitigates hallucination: rather than the model inventing a narrative, it is constrained to the factual graph traversal results (dates, record types, names) before it. In essence, the KG-RAG approach enabled the chatbot to function as a virtual research assistant, automatically collating archival evidence about John Howard's life events and summarizing them.

Quantitative/Aggregate Query: Another limitation of the simple RAG based ChatLoS version was the inability to handle aggregate questions (since LLMs are not reliable at arithmetic and the



chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://msa.maryland.gov/megafile/msa/speccol/sc5400/sc5496/domestic_traffic_ads/pdf/18290808ctpa6.pdf

msa.maryla... / **18290808ctpa6**

Sheriff' Sale.

Will be sold on aturday the 15th day of August next, a the Court house Door in Centreville,
Between the hours o 10 & 5 o'clock
of that day, fo Cash

One Negroman, named

JOHN HOWARD.

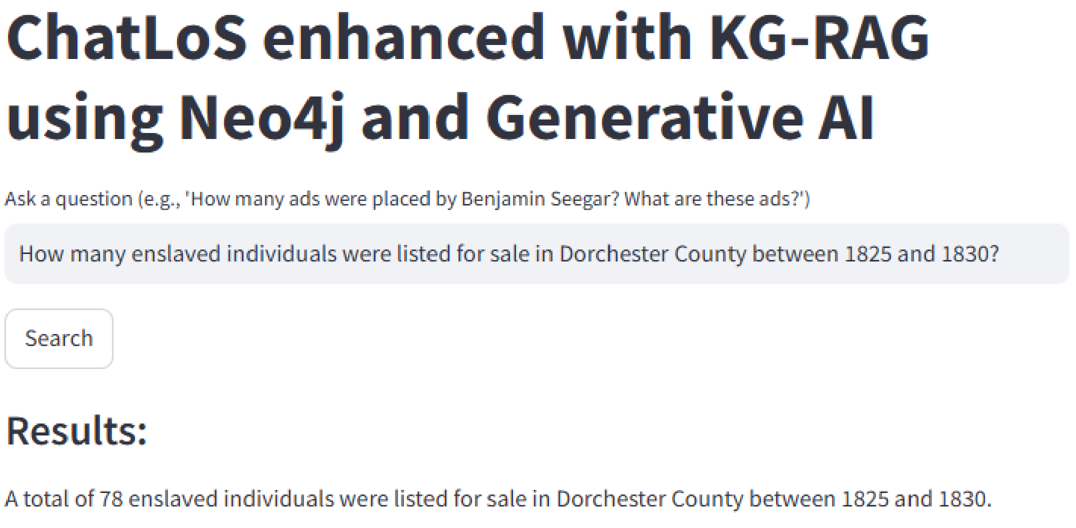
(*For a Term of yrs,*) the property of Stephen Boon, tan in Execution, by Virtue of a Writ Fi. Fa. to me directed, at the S of Wm. Reed, against said Boon.

Th. Aschom.

July 25th t *Sheriff.*

Figure 3: 1829 sale advertisement for John Howard

RAG approach only fetched a few documents at a time). With the KG and Cypher-driven retrieval, our system can answer questions involving counting or other simple analytics by delegating that task to the database. For example, we asked: *"How many enslaved individuals were listed for sale in Dorchester County between 1825 and 1830?"* – a question that requires filtering the DTA records by county and date, and counting results. The ChatLoS (KG-RAG) system produced a correct, succinct answer: "A total of 78 enslaved individuals were listed for sale in Dorchester County between 1825 and 1830." Figure 4 shows the interface's response. The user's query (in the input box) asks for the number of individuals in ads for a specific county and time range. The system responds with a factual sentence giving the count (*"78 enslaved individuals. . . "*). Under the hood, the system generated a Cypher query to count `DomesticTrafficAd` nodes with county = 'Dorchester' and date between 1825 and 1830, demonstrating the ability to perform database-style aggregation and then verbalize the result. This functionality was not possible in the initial RAG chatbot, which lacked direct data aggregation capability. The aggregate answer shows how integrating a database query capability addresses a practical user need. In the LoS archives, researchers often want statistics (e.g., how many ads in X county, how many people freed in Y year, etc.) which previously required manually exporting data or using tools like Tableau. Now, such queries can be asked in plain English. In our comparison demo, the MSA director saw that asking ChatLoS yielded the answer instantly, as opposed to manually applying filters in Tableau – he noted the ease of the natural language approach



ChatLoS enhanced with KG-RAG using Neo4j and Generative AI

Ask a question (e.g., 'How many ads were placed by Benjamin Seegar? What are these ads?')

How many enslaved individuals were listed for sale in Dorchester County between 1825 and 1830?

Results:

A total of 78 enslaved individuals were listed for sale in Dorchester County between 1825 and 1830.

Figure 4: KG-RAG ChatLoS Querying sale records by county and date

and how it *"handled open-ended questions"* without requiring step-by-step filtering. The system's ability to interpret the question, form a precise query, and return an exact number addresses a gap in the original system's functionality. Moreover, because the answer is grounded in an exact database result, it increases trust in its accuracy (the number "78" can be verified by running the same query in Neo4j or looking at the records, unlike an end-to-end neural answer which might be unsure). In the interview participant's words, the chatbot was *"more conversational and easier to understand"* than manually scanning database results. In summary, these examples illustrate that (i) the KG enables complex cross-record reasoning that was previously infeasible, (ii) integrating Cypher-based retrieval allows the chatbot to give precise answers to analytical queries, and (iii) the system can present information in a structured, source-linked manner, enhancing transparency. Next, we turn to the insights provided by the expert user who tried these features and shared his perspective.

5.2 User Feedback (Case Study Findings)

The interview with the LoS director yielded rich feedback on both the current system limitations and the new KG-RAG chatbot's potential. We summarize the key themes from the discussion:

Current System Pain Points: The participant described how researchers currently use the LoS online database and highlighted significant limitations. Notably, the inability to search across multiple datasets or criteria at once was a recurring frustration. For example, if a user wants to find records for a person across different record types, they must run separate searches in each dataset. He stated *"we can only do one [dataset] at a time... it would be great to be able to do multiple at the same time"*. Combining filters (like county and time period) is possible to an extent in the current system, but if a question spans dimensions (e.g., *"enslaved people in Anne Arundel*

County who later appear in Certificates of Freedom"), it's practically impossible with the existing interface. He also noted that some scholars resort to requesting entire datasets and then writing their own scripts or using Excel – an indicator that the current tools are not sufficiently flexible for advanced queries. This feedback established a baseline need that our system aims to meet.

Approachability of a Chatbot Interface: When introduced to the ChatLoS concept, the participant reacted positively to the idea of querying in natural language without needing database knowledge. He confirmed that the typical users (genealogists, students) are *non-technical* and currently don't require training to use the basic search interface, but that interface is limited. A conversational system could lower the barrier further by letting users ask questions in plain English, especially for those who *"don't know the underlying data structure"*. He found the chatbot format *"more approachable for general users"* and appreciated that it abstracts the complexity of the schema. This aligns with our design goal of democratizing access. Importantly, after seeing examples, he commented that ChatLoS was *"easier to understand... especially for non-technical users"* compared to a tool like Tableau. He envisioned this being useful for patrons who come with broad questions without knowing where to begin – the chatbot could handle the discovery aspect interactively.

Usefulness of KG Context: The participant was enthusiastic about the KG integration. Seeing the system trace an individual across records addressed a major *"fragmentation"* problem in the LoS ecosystem. He remarked on the *"ability to connect multiple datasets and provide comprehensive answers"* as a breakthrough, since currently even staff have to manually correlate information from separate sources. The example of John Howard's journey resonated strongly; he noted that the system essentially did what an expert researcher might do – link an ad to a manumission to a freedom certificate – but in seconds. This not only has user-facing value but could aid archivists themselves in tasks like verifying data linkages. Moreover, the participant had prior experience with data visualizations of LoS (e.g., network graphs of people) and said those were helpful in revealing patterns. The KG is like an underlying visualization that the chatbot taps into. He agreed that showing the intermediate connections (e.g., the chain of events for a person) *"adds credibility"* to the answer, because it mirrors how a human would justify a conclusion by citing sources. This feedback supports the idea that transparency via KG paths can improve trust.

Perceived Accuracy and Limitations: While impressed, the participant also discussed concerns and areas for improvement. One concern was the potential for varying responses or errors with nuanced queries. He observed during testing that slight rephrasing of a question sometimes led to different answers or the need for the bot to clarify, which could be frustrating if not handled. This is a known challenge with LLMs (non-determinism). For instance, if a question was too broad, the system might need to ask a follow-up or might give a partial answer. We explained how the system currently handles this (e.g., using the KG to narrow down possibilities, or prompting the user to clarify a name or date range). The participant suggested this is acceptable as long as the system guides the user. Another limitation discussed was *Name Recognition* – the LoS data often has inconsistencies (spelling variations of names, etc.). The user noted that genealogists might ask *"Find X person"* not realizing the name is spelled differently in records. The system's reliance on exact matches (unless a sophisticated alias mechanism is in place) could cause it to miss relevant results. This is a classic archival issue and he recommended exploring fuzzy name matching or integrating known variant name lists in the KG. We acknowledge this as a current limitation: the KG has to be populated with as many cross-references as possible, but it's not perfect.

Trust and Ethical Considerations: A significant discussion point was the trustworthiness of the AI's answers. The participant appreciated that the system is grounded in actual records and, when it showed sources or links, he felt more confident in the answer. He explicitly stated that seeing how the answer was derived (e.g., knowing it pulled from a verified database and not *"making things up"*) is crucial for him to trust using it in research. We showed how the chatbot could include references. He agreed this was a good approach and even suggested possibly surfacing the underlying query or graph path in an *"explain answer"* mode for power users – essentially making the black box more transparent.

User's Future Needs and Ideas: Looking ahead, the participant expressed interest in several expansions. One was incorporating more datasets (beyond the three we used) into the KG – for instance, the Slave Statistics or Runaway Ads collections that are also part of LoS. He imagined being able to query across all LoS datasets eventually, which would be extremely powerful for researchers. Another idea was to create different modes or interfaces for different user groups: for example, a simplified Q&A interface for the general public vs. a more data-centric analysis mode for scholars. This could possibly involve separate graphs or controlled vocabularies. This indicates broader applicability of our approach to other archival research projects. Finally, he was open to the idea of co-developing an evaluation rubric specifically for archival Q&A tools (we had asked if he would help define what a *"good answer"* means in this context – balancing accuracy, nuance, source evidence, etc.). He was receptive, which bodes well for our plan to carry out a community-based evaluation with multiple archivists in the future. In summary, the user feedback was enthusiastic about the concept (*"this could really make the information more accessible and engaging"* and validated the core features of KG-RAG ChatLoS as addressing real needs (multi-dataset search, ease of natural language, trust via source linking). The feedback also highlighted important considerations for refinement: handling name ambiguity, ensuring consistent behavior, and expanding content coverage. These insights feed directly into our discussion of the system's implications and the next steps for research.

A detailed comparison table on the evaluation of the performance of ChatLoS across its three versions vs the traditional data visualization tool as assessed based on the user study with the MSA's director is found in 3.

6. Limitations & Future Work

While the results from our KG-RAG system and expert user evaluation are promising, several limitations remain. The user study was limited to a single domain expert, whose feedback, while valuable, may not represent the broader user base, including students, genealogists, or general researchers. Future evaluations will need to include diverse participants to assess usability and usefulness more rigorously. We also acknowledge potential researcher bias due to our deep involvement in the system design and interpretation. The KG currently covers only three LoS datasets.

Table 3: Comparison of Capabilities Across ChatLoS versions and Tableau

Capability / Evaluation Criterion	**Tableau (traditional viz tool)**	**ChatLoS simple RAG**	**ChatLoS CSV-Agent**	**ChatLoS KG-RAG**
Natural-language question answering on a *single* dataset	X	✓	✓	✓
Interactive follow-up / conversational refinement	X	✓	✓	✓
Automatic semantic mapping of user terms to data columns	X	X	✓	✓
Aggregate counts & multi-attribute filters	✓	X	✓	✓
Cross-collection linking and reasoning	X	X	✓	✓
Multi-hop reasoning / trace life-paths of entities	X	X	X	✓
Built-in visualization quality (charts / maps)	✓	X	X	X
Explainability (surfacing code / query sent to engine)	✓	X	✓	✓
Typical response speed / user wait time	Fast	Slow	Slow	Slow
Zero-coding ease-of-use for non-technical staff	X	✓	✓	✓

Legend: ✓ = fully supported X = not natively supported

Although linkages between records were verified where possible, the possibility of missed or incorrect connections exists, particularly due to name variations. The KG schema, while effective for core use cases, simplifies real-world complexity and omits factors like family ties or mobility patterns. Expanding the graph will introduce challenges in both performance and LLM-driven query generation. Reliance on GPT-4o introduces non-determinism, in that, query generation can vary, and occasional errors in syntax or logic may cause failures. While our prompt engineering mitigates most issues, more robust error handling and refinement loops (e.g., agentic LLMs) are needed. Additionally, the system is limited to a basic interface without interactive graph exploration or visual traceability. Future work will address these gaps. We plan to conduct formal user evaluations (by involving communities of descendants, archival patrons and practitioners), expand the KG with additional LoS datasets and external references, address the named entity recognition issues, explore the possibilities in creating two interfaces according to external and internal facing audiences, create an archival evaluation rubric and explore hybrid retrieval with semantic search. We also intend to build agent-based query refinement and enhance the user interface for transparency, explainability and accessibility. These developments will help scale the system from a prototype to a deployable research tool.

7. Conclusion

To answer RQ1 posed for this project, we presented a KG-RAG ChatLoS enhanced chatbot conversational interface that unified three archival datasets from the LoS project and enables conversational exploration using Cypher-driven LLM queries. The KG encoded relationships across these datasets, allowing users to trace individual life paths and uncover insights beyond traditional keyword search. The LLM enabled natural language querying and generates structured, explainable responses. An expert user study highlighted the strengths of this approach, particularly its ease of use, historical accuracy, and ability to reveal connections that are difficult to surface in siloed databases. By grounding LLM responses in structured, verifiable relationships, the system addressed trust and explainability challenges often associated with generative AI in the humanities. While still early-stage, this work demonstrates how combining KGs with LLMs can create accessible, trustworthy AI systems for cultural heritage. Our next steps focus on broader evaluation, interface design, and cross-domain generalization. We believe this architecture can be adapted to other archival collections, supporting a new generation of explainable, user-centered tools for historical research and public memory.

8. Acknowledgment

We wish to acknowledge support from a grant from The Institute of Museum and Library Services (IMLS) Laura Bush 21st Century Librarian (LB21 program called: “Harnessing Generative AI to Support Exploration and Discovery in Library and Archival Collections.” This applied research project (funded in 2024) focused on the IMLS Goal of “Improving the ability of libraries and archives to provide broad access to and use of information and collections, and the IMLS Objective of “Supporting innovative approaches to digital collection management.” We are very grateful for this essential support and the selfless and dedicated staff at IMLS who promote innovation and “a nation where individuals and communities have access to museums and libraries to learn from and be inspired by the trusted information, ideas, and stories they contain about our diverse natural and cultural heritage.

References

[1] M. Caswell, *Toward a survivor-centered approach to records documenting human rights abuse: lessons from community archives*, Archival Science **14** (2014) 307.

[2] R. Marciano, *AFTERWORD: Towards a new Discipline of Computational Archival Science (CAS)*, Bielefeld University Press (2021).

[3] S. Patton, “Ai meets archives: The future of machine learning in cultural heritage.” CLIR (Council on Library and Information Resources) Blog, October 2024, 2024.

[4] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain et al., *Large language models: A survey*, 2024.

[5] Appian, “What is generative ai in simple terms?.” `https://appian.com/learn/topics/enterprise-ai/what-is-generative-ai`, 2024.

[6] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang et al., *A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions*, *ACM Transactions on Information Systems* **43** (2025) 1–55.

[7] E.M. Bender, T. Gebru, A. McMillan-Major and S. Shmitchell, *On the dangers of stochastic parrots: Can language models be too big?*, in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 610–623, ACM, 2021.

[8] J. Larson and S. Truitt, “GraphRAG: Unlocking LLM Discovery on Narrative Private Data.” `https://www.microsoft.com/en-us/research/blog/graphrag-unlocking-llm-discovery-on-narrative-private-data/`, February, 2024.

[9] R.K. Gnanasekaran and R. Marciano, *Leveraging OpenAI’s LLMs and Cloud-based Learning-as-a-Service (LaaS) Solutions to Create Culturally Rich Conversational AI Chatbot: ChatLoS - A Study Using the Legacy of Slavery Dataset*, in *International Symposium on Grids & Clouds (ISGC) 2024*, p. 1, Oct., 2024.

[10] L. Xu, L. Lu, M. Liu, C. Song and L. Wu, *Nanjing yunjin intelligent question-answering system based on knowledge graphs and retrieval augmented generation technology*, *npj Heritage Science* **12** (2024) .

[11] A. Inbasekaran, R.K. Gnanasekaran and R. Marciano, *Using transfer learning to contextually optimize optical character recognition (OCR) output and perform new feature extraction on a digitized cultural and historical dataset*, in *2021 IEEE International Conference on Big Data (Big Data)*, pp. 2224–2230, DOI.

[12] L.A. Perine, R.K. Gnanasekaran, P. Nicholas, A. Hill and R. Marciano, *Computational treatments to recover erased heritage: A legacy of slavery case study (ct-los)*, in *2020 IEEE International Conference on Big Data (Big Data)*, pp. 1894–1903, 2020, DOI.

[13] R.K. Gnanasekaran and R. Marciano, *Piloting data science learning platforms through the development of cloud-based interactive digital computational notebooks*, .

[14] R.K. Gnanasekaran, L. Perine, M. Conrad and R. Marciano, *Model selection for heritage-ai: Evaluating llms for contextual data analysis of maryland’s domestic traffic ads (1824–1864)*, in *2024 IEEE International Conference on Big Data (BigData)*, pp. 2419–2430, IEEE, 2024.

[15] J. Greenberg, *Ai-ready data: Navigating the dynamic frontier of metadata and ontologies - a workshop summary*, 2024.

[16] C.S.G. Khoo, E.A.L. Tan, S. Ng, C. Chan, M. Stanley-Baker and W.-N. Cheng, *Knowledge graph visualization interface for digital heritage collections: Design issues and recommendations*, *Information Technology and Libraries* **43** (2024) .

[17] *How to query a knowledge graph with llms using grag*, *Towards Data Science* (2024) .

[18] *How to convert any text into a graph of concepts: A method to convert any text corpus into a knowledge graph using mistral 7b*, *Towards Data Science* (2023) .

[19] *Knowledge graph, ai services and next generation instrumentation for research and development in social sciences and humanities*, 2024.

[20] *Llms in each stage of building a graph rag chatbot: A case study*, *KuzuDB Blog* (2024) .

[21] *Llm knowledge graph builder - first release of 2025*, *Medium* (2025) .

[22] H.D. Nguyen, T.-H.A. Nguyen and T.B. Nguyen, “A proposed large language model-based smart search for archive system.” arXiv preprint arXiv:2501.07024, 2025.