

## Harnessing Generative AI to Support Exploration and Discovery in Library and Archival Collections

### 1. Introduction

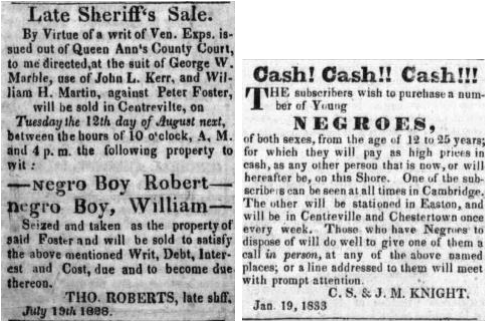
The University of Maryland iSchool, in collaboration with the Maryland State Archives Legacy of Slavery Program and the Kennard African American Cultural Heritage Center in Queen Anne’s County, Maryland, is submitting a 2-year applied research proposal to the IMLS National Leadership Grant for Libraries (NLG-L) program. We propose an innovative research project that harnesses the potential of generative AI to support exploration and discovery in library and archival collections. The project is designed to address critical needs within the library and archives fields, aligning with NLG-L Goal 3 & Objective 3.2: ‘*Improve the ability of libraries and archives to provide broad access to and use of information and collections*’ & ‘*Support innovative approaches to digital collection management*’.

### 2. Project Justification

We see great potential for incorporating generative AI technologies such as ChatGPT into future library and archival services. ChatGPT is an example of conversational AI underpinned by a generative pretrained transformer (GPT) model. A GPT model is an advanced Large Language Model (LLM) trained extensively on diverse textual data, including books and articles. This training equips LLMs with the capability to generate natural-sounding, human-like text, rendering them highly valuable for tasks such as answering questions and summarizing context.

Our proposed initiative builds upon recent results of an IMLS-funded effort, “*Piloting an Online Collaborative Network*” project (grant [RE-246334-OLS-20](#)). This pilot study deployed a subset of the Maryland State Archives (MSA) Domestic Traffic Ads (DTA) collection to evaluate the effectiveness of GPT models.<sup>1</sup> The DTA collection encompasses newspaper advertisements from 1824 to 1864, shedding light on the abhorrent practices of chattel slavery—where buyers and sellers engaged in transactions involving human beings for social and domestic purposes.

**Table 1.** DTA advertisement examples, Types of GPT-powered AI bot queries, and Lessons learned from the pilot.

DTA advertisement samples for a sale ad and a purchase ad	GPT-powered AI bot queries using a subset of 764 of the DTA ads	Lessons learned from the pilot
	<ul style="list-style-type: none"> <li>● How many ads are in the dataset?</li> <li>● How many ads were placed in Queen Anne’s County, Maryland?</li> <li>● Were any ads placed on Christmas Day?</li> <li>● Can you report ad counts for both public auctions and public sales?</li> <li>● Give me ad counts by year between 1830 and 1835?</li> <li>● Do any ads mention Philip Benson?</li> <li>● Can you summarize these ads?</li> <li>● Can you find ads that relate to this one?</li> </ul>	<ul style="list-style-type: none"> <li>● DTA content was not suitable for use with generative AI and required significant pre-processing and curation</li> <li>● GPT models need to understand the context of the DTA collection and be further trained.</li> <li>● GPT models seem to be able to duplicate some of the level of analysis seen in non-AI exploratory case studies.</li> <li>● GPT models have parameters that allow for some control of trustworthiness.</li> </ul>

LLMs, such as GPT, have the potential to empower libraries and archives to broaden access to their collections and support plain English text interfaces for querying collections and generating text responses, aligning with NLG-L Goal 3. Furthermore, they offer innovative avenues for interaction with digital collections, in line with NLG-L Objective 3.2. Three fundamental questions guide our research:

- (RQ1) *How should we further curate library and archival collections to make them GPT-model ready?*
- (RQ2) *How can we compare GPT models with traditional, non-AI exploratory data analysis models?*
- (RQ3) *How can we incorporate sociotechnical assessments of trustworthiness and detect bias in GPT models?*

By addressing these research questions, our project endeavors to push the boundaries of library and archival services through the creative application of generative AI. In doing so, we aim to strengthen the ability of libraries and archives to serve the public and research communities by unlocking new dimensions of access, discovery, and understanding within their

<sup>1</sup> UCL Press/AEOLIAN edited collection: *Artificial Intelligence for Cultural Heritage Organisations across the Atlantic*. “Conversing with the Past: Re-examining the Legacy of Slavery in Domestic Traffic Newspaper Advertisements with OpenAI’s GPT3 LLM”. **Rajesh Kumar Gnanasekaran, Christopher E. Haley, and Richard Marciano**. To be published in 2024.

collections. Through this NLG-L grant, we seek to create a lasting impact by fostering innovation, inclusivity, and excellence in the library and archival professions.

### 3. Project Work Plan

The research team comprises members of the Advanced Information Collaboratory (AIC) at the UMD iSchool, dedicated to exploring the opportunities and challenges of “disruptive technologies” for archives and records management. We aim to leverage cutting-edge technologies to unlock hidden information within extensive record collections. All members of the team participated in a [2-day Oct. 2019 datathon](#) held at the MSA, with participation from King’s College London and the UK National Archives. The datathon focused on computational treatments of the Legacy of Slavery collections, encompassing runaway ads, certificates of freedom, and manumission records. Four leading AIC experts will direct the research tasks (see Table 2). The project has three phases: Phase 1 (Aug. ‘24 - Mar.. ‘25) tackling RQ1, Phase 2 (Jan. - Dec. ‘25) tackling RQ2, and Phase 3 (Oct. 25. - Jul. ‘26) tackling RQ3.

**Table 2.** Research questions, leads, and approaches

	Research Leads	Research Approach
<b>RQ1</b>	<ul style="list-style-type: none"> <li>● <b>Richard Marciano</b>, Professor, Univ. of Maryland iSchool.</li> <li>● <b>Rajesh Kumar Gnanasekaran</b>, Doctoral Candidate, iSchool, Univ. of Maryland, campus IT director.</li> </ul>	<ul style="list-style-type: none"> <li>● Research collection curation techniques to support LLM work and refine LLM models to understand a collection’s context.</li> </ul>
<b>RQ2</b>	<ul style="list-style-type: none"> <li>● <b>Mark Conrad</b>, Digital Archive Research, Digital Archives Education (formerly with the US National Archives).</li> </ul>	<ul style="list-style-type: none"> <li>● Conduct case studies using traditional, non-AI approaches for comparison.</li> </ul>
<b>RQ3</b>	<ul style="list-style-type: none"> <li>● <b>Lori Perine</b>, AIC Research Fellow, and Associate Researcher in Trustworthy AI at National Institute of Standards and Technology (NIST). See: <a href="#">bias</a> &amp; <a href="#">playbook</a>.</li> </ul>	<ul style="list-style-type: none"> <li>● Develop a sociotechnical framework for bias mitigation and community engagement in AI system design and assessment.</li> </ul>

### 4. Diversity Plan

We are committed to fostering diversity and inclusivity in our project. To achieve this, we will collaborate closely with Christopher Haley, the Director of Research for the Study of the Legacy of Slavery (LoS) at the Maryland State Archives. Mr. Haley serves on the Maryland Lynching Commission for Truth and Reconciliation and is associated with the Kunta Kinte-Alex Haley Foundation. The LoS project has yielded a web-accessible database comprising over 400,000 pieces of information across 17 diverse and interconnected collections. These collections (including Certificates of Freedom (CoF), Domestic Traffic Ads (DTA), Inventories, Manumissions (M), MD Penitentiary Records, Runaway Ads (RA), Slave Jails, and Slave Schedules) are invaluable for recognizing and identifying previously unknown African American residents of Maryland. Additionally, we will collaborate with the Kennard African American Cultural Heritage Center in Centreville, Queen Anne’s County, Maryland for community engagement in assessing trustworthiness and bias in the outcomes of generative AI interfaces that access the DTA collection. The DTA and other LoS collections include many records associated with enslaved and formerly enslaved people in Queen Anne’s County.

### 5. Project Results

The DTA collection is one of 17 distinctive LoS collections housed by the MSA. The MSA’s DTA is a rare and unique collection of digitized documentation of freed and emancipated Black populations in the United States. Grounding this applied research grant around the DTA, we expect to develop generalizable approaches across more commonly available collections for access and exploration, and to offer guidance to using generative AI with library and archival collections at large. Our AI models and interfaces will be shared through GitHub and Jupyter Notebook case studies in order to be reused by other institutions. Furthermore, our project has the potential to directly contribute to the training of library and archives professionals through the creation of a course on the application of LLMs to libraries and archives. Results will be disseminated through ALA, SAA, NAGARA, the UMD iSchool DCIP professional certificate in digital curation (taught by Conrad and Marciano), graduate courses in the UMD iSchool and partner iSchools across the US through the IMLS-funded TALENT Network ([RE-252287-OLS-22](#)).

### 6. Budget Summary

The University of Maryland iSchool requests \$194,000 over a two-year period (Aug. 1, 2024 to Jul. 31, 2026) as follows: \$46K in salaries and wages, \$3K in fringe benefits, \$20K in travel, \$10K in computing costs, \$35K in research leads/consultants, \$10K in participant stipends (total direct costs of \$124K), with \$70K in indirect costs (56% indirect).