

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

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 for \$194,229. We propose an innovative applied 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.’*

While this proposal is motivated by a 2023 pilot study (described in the next section) where we used [ChatGPT](#) (to experiment with the formulation of plain English queries on a subset of the Maryland State Archives (MSA) Domestic Traffic Ads (DTA) collection, we propose to evaluate the use of other open-source GPT models (see section 1 for definition) in the first phase of this project. The use of ChatGPT may carry a significant disadvantage: the potential to transmit sensitive and potentially offensive content back to OpenAI, with the risk of it being incorporated into their training data. The decontextualized information in the DTA advertisements can lead to ChatGPT outputs containing inflammatory and dehumanizing content learned from historical materials where racial and societal norms were different from contemporary norms. Fortunately, now several GPT models shield content from being reincorporated into the training data. These include (at this moment!) Llama2 open source, Microsoft Azure OpenAI GPT models, Amazon Web Services (AWS)’s Anthropic Claude 2.1, and Google’s Gemini. We will re-evaluate these options at the start of the project to select the most appropriate solution(s).

### 1. Project Justification

We see great potential for incorporating language-based generative AI technologies into future library and archival services. One commonly known example is ChatGPT, a conversational AI underpinned by a Generative Pre-trained Transformer (GPT) model developed by OpenAI. GPT models are a subset of advanced Large Language Models (LLMs)<sup>1</sup>, which are trained extensively on diverse textual data, including books and articles. This training equips GPT models with the capability to generate natural-sounding, human-like text responses, rendering them highly valuable for tasks such as answering questions and summarizing context. A recent example of this is the open-source WARC-GPT project<sup>2</sup> from Harvard’s Library Innovation Lab, which was developed to interrogate web archiving content in the form of WARC files (a file format that is a revision and generalization of the ARC format used by the Internet Archive to store information blocks harvested by web crawlers), by asking specific questions in natural language rather than relying on keyword searches and metadata filters.

We feel a sense of urgency to conduct this work as it emphasizes the inclusion of library and archival holdings as data for AI tools. Research around GPT models has exploded over the last year, with dizzying speed, new solutions, new companies, and a push for their responsible use. A recent article<sup>3</sup> makes a compelling case for the dangers of generative AI even obscuring the cultural record. Thus, anchoring our study with a unique and sensitive historical collection, informed by relevant stakeholder input, should provide valuable feedback to librarians and archivists in the eye of the ongoing AI storm. Library and archival holdings must inform this conversation.

Our proposed initiative builds upon recent results of an IMLS-funded effort, the *“Piloting Network”* project (grant [RE-246334-OLS-20](#)) where we evaluated the effectiveness of GPT models<sup>4</sup>, by conducting a 2023 pilot study using ChatGPT queries with a subset of the MSA Domestic Traffic Ads (DTA) collection. The DTA collection encompasses newspaper advertisements from 1824 to 1864, which shed light on the abhorrent practices of chattel slavery—where buyers and sellers engaged in transactions involving human beings. (see Table 1):

<sup>1</sup> What are LLMs, and how are they used in generative AI?, Lucas Mearian, ComputerWorld, Feb. 7, 2024. See: . See:

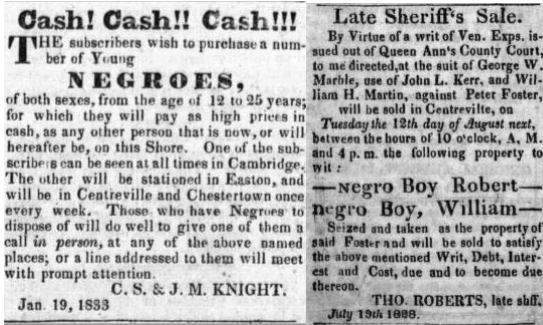
<https://www.computerworld.com/article/3697649/what-are-large-language-models-and-how-are-they-used-in-generative-ai.html>.

<sup>2</sup> WARC-GPT: An Open-Source Tool for Exploring Web Archives Using AI, Posted by M. Cargnelutti, K. Mukk, and C. Stanton, February 12, 2024. See: <https://lil.law.harvard.edu/blog/2024/02/12/warc-gpt-an-open-source-tool-for-exploring-web-archives-with-ai/>.

<sup>3</sup> “A.I. Is Coming for the Past, Too”, New York Times, OPINION GUEST ESSAY: by J. N. Shapiro and C. Mattmann, <https://www.nytimes.com/2024/01/28/opinion/ai-history-deepfake-watermark.html>, Jan. 28, 2024.

<sup>4</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”. **R. Kumar Gnanasekaran, C. E. Haley, and R. Marciano**. To be published in 2024.

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

DTA advertisement samples for a purchase ad and a sale ad	GPT-powered AI bot queries with a subset of 764 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>● Can you 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 AI-ready and needed significant pre-processing to lend itself to generative AI.</li> <li>● GPT models can understand natural language statements to automatically choose the columns needed for aggregate data analysis.</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>

Customized or adapted GPT models have the potential to empower libraries and archives to broaden access to their collections and support plain English text user interfaces for querying collections and generating text responses, aligning with NLG-L Goal 3. Furthermore, they offer innovative avenues for interaction with digital collections, aligning with NLG-L Objective 3.2. Three fundamental questions emerged from our 2023 pilot study and guided our research in moving forward:

- **(RQ1)** *How should we further curate library and archival collections to make them AI-ready?*
- **(RQ2)** *How can we compare GPT models with traditional, non-AI exploratory data analysis models?*
- **(RQ3)** *How can we incorporate socio-technical considerations<sup>5</sup> to promote trustworthiness and mitigate potential bias arising from the use of GPT models with library and archival collections?*

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 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.

We view **RQ1** as the enabling research question that justifies our research with library and archival collections. Fortunately, since we submitted our pre-proposal, studies on the nature and form of “AI-ready data” are increasingly being conducted. The research leads of this IMLS proposal were invited to contribute to an April 15-16, 2024, NSF workshop on this topic. AI-ready data refers to the high-quality and well-prepared data that is optimized for use in AI applications: “AI-ready data increasingly encompasses the inclusion of metadata and ontologies to enhance the value and usability of data to help data scientists, researchers, and AI systems understand, interpret, and apply appropriate algorithms and models for analysis.”<sup>6</sup> Advances in this area are also expected to help support FAIR (Findable, Accessible, Interoperable, and Reusable) principles and reproducible computational research (RCR). Ongoing work by the NOAA Center for AI<sup>7</sup> has even proposed an AI-readiness matrix that maps data quality, data access, and documentation in terms of levels of readiness described as Level

<sup>5</sup> NIST Special Publication 1270, “Towards a Standard for Identifying and Managing Bias in Artificial Intelligence”. R. Schwartz, A. Vassilev, K. Greene, L. Perine, A. Burt, P. Hall. March 2022. See: <https://doi.org/10.6028/NIST.SP.1270>.

<sup>6</sup> AI-Ready Data: Navigating the Dynamic Frontier of Metadata and Ontologies, ID4: Institute of Data Driven Dynamical Design. Hosted by the Metadata Research Center, College of Computing & Informatics, Drexel University, April 15-16, 2024. See: <https://mrc.cci.drexel.edu/2024/02/23/ai-ready-data-navigating-the-dynamic-frontier-of-metadata-and-ontologies/>.

<sup>7</sup> NOAA Workshop on Leveraging AI in Environmental Sciences, Oct. 2020. Tyler Christensen et al, 10/22/2020. See: [https://www.star.nesdis.noaa.gov/star/documents/meetings/2020AI/presentations/202010/20201022\\_Christensen.pdf](https://www.star.nesdis.noaa.gov/star/documents/meetings/2020AI/presentations/202010/20201022_Christensen.pdf).

0/Not AI-Ready, Level 1/Minimal, Level 2/Intermediate, and Level 3/Optimal. Library and archival collections need to be represented in these forums.

Our additional research questions address two important dimensions contributing to AI-readiness: trustworthiness and relevance of GPT model outputs. For library and archival collections, validity, reliability, and reproducibility contribute to the trustworthiness of the model outputs, while contextual interpretability is key to the relevance of the output. It is important to address these dimensions to ensure that the use of computational methods broadly and GPT models specifically are implemented in ways that reveal rather than obscure cultural and historical records. Using conventional data science techniques, we interrogate the DTA collection using non-AI data science techniques to establish a “ground truth” in information that can be mined from the collection. This is the purpose of **RQ2**, which provides a benchmark for evaluating the accuracy and robustness of results using GPT models and assessing what innovative uses can be enabled via the GPT model. We incorporate socio-technical dimensions of trustworthiness and relevance with **RQ3**. Using input from stakeholder engagement, we intend to operationalize user norms and expectations in data and model preparation; evaluate the cultural and contextual relevance of GPT model outputs; and assess the validity of GPT outputs relative to experts’ and community domain knowledge.

## 2. Project Work Plan

The core research team comprises members of the [Advanced Information Collaboratory](#) at the UMD iSchool, dedicated to a research agenda that is highly relevant to this grant proposal: (1) Exploring the opportunities and challenges of "disruptive technologies" for archives and records management (digital curation, machine learning, AI, etc.), (2) Leveraging the latest technologies to unlock the hidden information in massive stores of records, (3) Pursuing multidisciplinary collaborations to share relevant knowledge across domains, (4) Training current and future generations of information professionals to think computationally and rapidly adopt new technologies to meet their increasingly large and complex workloads, and (5) Promoting ethical information access and use.

Our core four-member research team will work with Christopher Haley from the MSA. All five participated in a 2-day [Oct. 2019 datathon](#) held at the MSA, with collaborators from King’s College London and the UK National Archives. This event was funded under an Arts and Humanities Research Council one-year International Research Networking grant for UK-US Collaborations in Digital Scholarship in Cultural Institutions. Our emphasis at the time was on demonstrating a Computational Archival Science (CAS) network to explore the application of computational methods to contextualize records within archival collections. The October 2019 datathon focused on traditional non-AI computational treatments of the LoS collections, encompassing runaway ads, certificates of freedom, and manumission records. Our proposed Applied Research project builds on this earlier work by formulating the following research questions (see Table 2):

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

Research Questions	Research Leads	Research Methods
<b>(RQ1)</b> How should we further curate library and archival collections to make them AI-ready?	<ul style="list-style-type: none"> <li>● <b>PI: Richard Marciano:</b> Professor, U. of Maryland iSchool.</li> <li>● <b>Rajesh Kumar Gnanasekaran:</b> Doctoral Candidate, iSchool, U. of Maryland, campus IT director.</li> </ul>	Research collection curation techniques and AI-readiness to support generative AI work and refine GPT models to understand a collection’s context.
<b>(RQ2)</b> How can we compare GPT models with traditional, non-AI exploratory data analysis models?	<ul style="list-style-type: none"> <li>● <b>Mark Conrad:</b> Digital Archivist, AI-Collaboratory (formerly with the U.S. National Archives).</li> </ul>	Interrogate the DTA dataset using traditional, non-AI approaches for comparison with GPT model results.
<b>(RQ3)</b> How can we incorporate socio-technical considerations to promote trustworthiness and mitigate potential bias arising from the use of GPT models with library and archival collections?	<ul style="list-style-type: none"> <li>● <b>Lori Perine:</b> AI-Collaboratory Research Fellow &amp; former Associate Researcher in Trustworthy AI at the National Institute of Standards and Technology (NIST). See: <a href="#">bias</a> &amp; <a href="#">playbook</a>.</li> </ul>	Engage domain experts and community members in GPT model design and evaluation via surveys and focus groups; and incorporate their contextual input for data preparation, prompt engineering, and model training.

Our work plan actively engages with collection content experts at the MSA in Annapolis, Maryland, and a focus group of local cultural and historical experts and community members at the Kennard African American Cultural Heritage Center in Centreville, Queen Anne’s County, Maryland (located on the Eastern Shore, where many of the DTA Collection ads originate). This stakeholder-driven approach is at the heart of how we propose to incorporate socio-cultural considerations into the design and evaluation of using a GPT model to explore the DTA collection. (RQ3). Input from expert and

community users will inform data preparation for AI readiness and decisions in model engineering and tuning. In other words, the exploration and discovery in library and archival collections we seek to support in this Applied Research grant will be tested and validated through community engagement interactions. We structure our work plan into 6 phases and around two focus group working meetings (Phases 3 & 5). (see Table 3):

**Table 3.** Timeline for the 6 major phases of the project.

Aug. 1, 2024 – Jul. 31, 2026	2024					2025												2026							
	A	S	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D	J	F	M	A	M	J	J	
Phase 1: GPT Selection & Focus Group Input for Model Design	█	█	█	█	█	█																			
Phase 2: Data Preparation & Model Training				█	█	█	█	█																	
* Phase 3: Focus Group Working Meeting I *									█	█	█														
Phase 4: GPT Performance Analysis & Re-tuning												█	█	█	█	█	█								
* Phase 5: Focus Group Working Meeting II *																	█	█							
Phase 6: Wrap-up																					█	█	█	█	█

We propose the following tasks under each of these 6 phases:

**1. Phase 1: GPT Selection & Focus Group Input for Model Design (August 2024 - January 2025)**

o **Task 1.1: Model Selection.** (RQ1)

With the advent of the GPTs, several relevant GPTs are available for this project. We will perform a detailed evaluation process to select the best one with the steps below.

- Preliminary Research and Shortlisting:
  - o Survey the Landscape: Begin with a review of the latest GPTs available, focusing on those with strong privacy controls designed to handle sensitive data responsibly.
  - o Shortlist Candidates: Based on initial research, create a shortlist of GPTs that promise content shielding capabilities. This would likely include models like Llama2, Microsoft Azure's GPT models, AWS's Anthropic Claude 2.1, and Google's Gemini, among others that may emerge.
- Proposed Evaluation Criteria:
  - o Sensitivity to Historical Context: The model must handle historical data sensitively, understanding the context without perpetuating biases.
  - o Privacy and Security: Evaluate each model's capabilities to ensure that the data fed into it remains confidential and is not used to train other models without permission.
  - o Customizability: The ability of the model to be updated with "grounded" domain-specific knowledge for specific tasks like understanding the historical context of the DTA dataset and generating responses that are accurate and respectful of the data's nature.
- Final Selection:
  - o Select the Optimal Model: Choose the model that excels in handling sensitive data, is customizable to the project's needs, and maintains privacy and security standards. A detailed evaluation report of GPT models will be created, emphasizing those with robust privacy controls and sensitivity to the historical context. This report will guide the AI framework's development, ensuring it aligns with the project's ethical and operational standards.

We expect the landscape to be very different when we start the project! We will select the optimal model after testing candidates.

o **Task 1.2: Designing a User-centered Conversational Chatbot.** (RQ3)

An integral part of this project is a well-designed, user-centered conversational chatbot interface. We will follow the steps below to design and implement the chatbot interface.

- Design a user-centered chatbot interface, focusing on user-friendly design principles to accommodate a wide range of users, from researchers to the general public (see Fig. 1).
- Implement the chatbot using a suitable programming framework, integrating the trained GPT to process and respond to user-specific queries.



- Ensure the chatbot can handle natural language inputs, provide informative responses, and guide users to discover insights about the dataset.
- Incorporate an interactive feedback mechanism with "Thumbs Up" and "Thumbs Down" icons for users to rate responses and a prompt for textual feedback for both responses (up to 200 characters). Below is an example of one such instance where a "Thumbs Down" is chosen. This feature is crucial for capturing specific user feedback for continuous improvement (see Fig. 2).

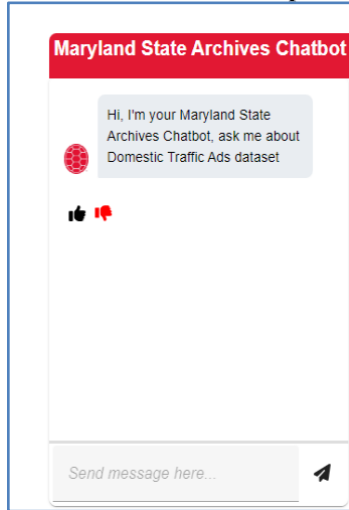


Fig. 1: Chatbot mockup

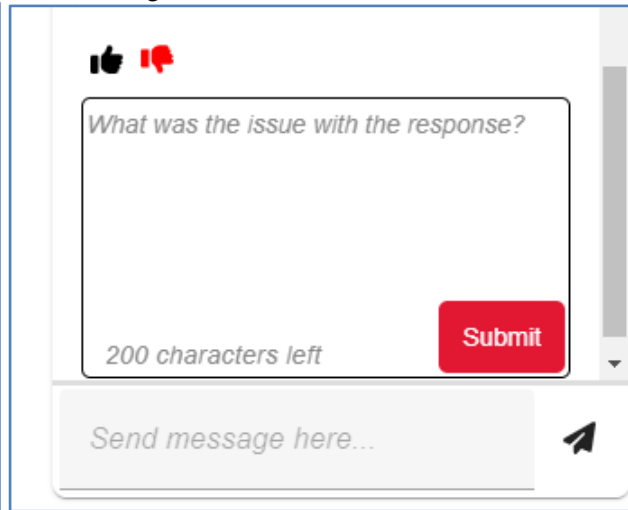


Fig. 2: Chatbot mockup

○ **Task 1.3: Focus Group Selection and Design Survey.** (RQ3)

With the assistance of Kennard African American Cultural Heritage Center and Maryland State Archives, we will define potential use cases for accessing the DTA Collection, such as historical research, K-12 education, higher education, look-up services in libraries or archives, interactive cultural heritage displays in museums, personal and professional genealogy, etc. Based on this information, we will collaborate with the Kennard Center to recruit a focus group of experts and community members representing the following user roles of the technology: researcher, historian, preservationist, teacher, student, and the general public. The focus group will have up to 15 members and will be engaged during Phases 1, 3, and 5. In this current phase 1, we plan to:

- Introduce the focus group to the DTA Collection and an outline of the common functionality related to the GPT models via a webinar.
- Capture initial input about expectations regarding usability and benefits of engaging the DTA collection via a GPT model and establish baseline concerns about risks or vulnerabilities (e.g. racial or gender bias, historical or cultural inaccuracies, privacy, etc.). Methods used include an asynchronous user design survey and follow-up interviews.
- Incorporate focus group expectations and concerns as metadata, features, or functionality to be used in the initial data preparation and/or model training. This step is articulated further in Phase 2. This type of operationalization also will serve as a baseline benchmark for evaluating system performance after Focus Group Working Meetings 1 and 2.

**2. Phase 2: – Data Preparation & Model Training (November 2024 - March 2025)**

○ **Task 2.1: Prepare Full DTA Collection for AI-Readiness.** (RQ1)

Our earlier 2023 pilot used a subset of the DTA collection of 764 records from nearly 3,000 total records (roughly 25%). We will now curate the entire dataset. In our pilot, we also augmented the dataset with six additional fields: (1) Terms of Service, (2) Trade Reason, (3) Features, (4) Terms of Sale, (5) Owner, and (6) the complete Optical Character Recognition (OCR) text of each DTA ad image. This initial transformed dataset was shown to support effective GPT-powered AI bot queries (see Table 1).

○ **Task 2.2: Incorporate User Input.** (RQ3)

During Phase 1, we will collect information from the focus group regarding expectations and risk concerns for using a GPT model to access the DTA collection. In Phase 2, we will investigate ways to operationalize that input, based on user role, to incorporate it into the dataset and/or make choices in GPT model design and training. Our approaches will include:

- Identifying indicators or proxies to represent expectations and/or risks specific to one or more user roles as metadata to be encoded via adding columns to the dataset.

- Engineering prompts to reflect the variability in syntax, contextual knowledge, and domain expertise among user roles and use cases.
- Training the GPT model to return results in a non-biased tone and using syntax consistent with contemporary norms for use cases that do not require historical accuracy. The objective would be to minimize amplifying past cultural biases while conveying relevant information.
- **Task 2.3: Generate non-AI Ground Truth of DTA Aggregated Data.** (RQ2)  
Anytime we update the GPT model, we will test the updates by comparing it with the results through a case study using a traditional, non-AI approach. This will be done for comparison and baseline purposes and will involve data science approaches. GPT models hold the promise of performing a statistical analysis without having to write a single line of code. While our project aims at conducting such advanced data analysis using plain English and more intuitive interfaces, we also generate a non-AI data analysis solution, so that we can assess the accuracy of the GPT responses.
  - Define Ground Truth Metrics:
    - Identify Key Metrics: Determine which metrics will be used to evaluate the DTA dataset. This could include counts, trends, patterns, and other statistical measures relevant to the historical data.
    - Set Evaluation Criteria: Establish criteria for assessing the accuracy, completeness, and relevance of the data analysis.
  - Traditional Data Analysis:
    - We choose to work with Jupyter Notebooks (there are many options though): Employ Jupyter Notebooks to write and execute Python code for data analysis. This involves creating scripts that can process and analyze the dataset without AI.
    - Statistical Analysis: Perform statistical analysis to extract the defined metrics from the dataset. This could include frequency counts, time series analysis, and correlation studies among various dataset dimensions.
  - Visualization:
    - Create Visualizations: Develop charts, graphs, and maps to visually represent the analysis findings. Visualizations help in understanding trends, patterns, and anomalies within the dataset.
    - Document Insights: Annotate visualizations with insights and observations that explain the significance of the findings.
  - Comparative Baseline Establishment:
    - Compile Ground Truth Data: Aggregate the results of the traditional data analysis into a comprehensive ground truth dataset.
    - Prepare for Comparison: Structure the ground truth data to be directly compared to the outputs generated by the GPTs. This may involve summarizing key findings, metrics, and visualizations in a comparative framework.
  - Review and Validation:
    - Peer Review: Conduct an internal or external peer review of the ground truth findings to validate the accuracy and reliability of the analysis.
- **Task 2.4: Prepare the GPTs for DTA Conversational Chatbot.** (RQ1, RQ3)
  - Integration of AI-Readiness and User Input
    - Based on Task 2.1 and 2.2, add metadata to the dataset that reflects user roles, expectations, and potential risk concerns. This involves adding new columns to the data that encode these aspects, ensuring the AI models can understand and use this information when generating responses.
  - Annotation and Augmentation for Contextual Clarity
    - Detailed Tagging and Highlighting: Extend the tagging process to include user-defined metadata and expectations, ensuring tags reflect the nuanced understanding required for different user roles and use cases.
    - Operationalize User Input: Utilize the insights from Phase 1 to identify and incorporate indicators or proxies for inclusion as data or metadata; guide the annotation process, ensuring data is marked for historical accuracy, sensitivity to contemporary norms, and the minimization of bias; and adjusting sensitivity thresholds when retuning the model.
  - Preparing the GPTs for querying the DTA dataset
    - The goal of this task is to prepare the GPTs to be compared in two distinct ways: one using the Retrieval Augmented Generation ([RAG](#)) method to limit the GPT's knowledge to only the context retrieved based

on the question asked, and another to use the entire dataset in the context memory for performing data aggregation analysis.

- In simple terms, the RAG method splits the entire DTA dataset into subsets or chunks of related data and stores them in a special database. This database can then be quickly searched in response to specific questions by users, thereby reducing the need to store the entire dataset in memory and increasing efficiency in the case of large datasets. This tuning is especially beneficial for questions specific to certain individuals, events, or activities. When asked a question, this is like preparing a quick reference guide for the AI. After Task 2.2, the subsets of the dataset are now enriched with user expectations and risk annotations, tailored for quick retrieval by the RAG-enhanced GPT models.
- The second tuning method uses the full DTA dataset in Contextual Memory for each conversation to utilize all available information, including user role metadata. This enables the GPT to perform data aggregation analysis on the entire dataset and respond with answers accordingly. Ensure the entire dataset, including user metadata and annotations for non-bias, is organized for the model that will use it, aiming for deep, comprehensive analysis.

### 3. Phase 3: Focus Group Working Meeting I. (March-May 2025)

We intend to convene the focus group in person at Kennard for a daylong, hands-on working meeting in May 2025. The objective of the working meeting will be to engage the focus group in exploring and evaluating the usability, contextual accuracy, cultural bias, and use case relevance of the chatbot. Usability will address the user experience with the chatbot interface. Contextual accuracy refers to user evaluation of the accuracy of chatbot output, relative to the user's expectations for retrieving and using information from the collection. Users will assess the chatbot for cultural bias and use case relevance. The working meeting will include both structured and unstructured exploration.

- **Task 3.1: Structured Exploration.** (RQ2, RQ3)

During the facilitated structured exploration, the research team will guide the focus group through a set of tasks and questions to explore the DTA collection using the chatbot. Some of the tasks will be those used for the model training and tuning in Phase 2, to provide a comparison with the non-AI ground truth and the initial baseline performance metrics established during model training. Other tasks and questions will be relevant to the use cases represented by the focus group. Individual evaluation and assessment of chatbot usability, contextual accuracy, cultural bias, and use case relevance will be gathered via real-time, live feedback polls. This will allow us to record and capture instant impressions, with collective responses available for display in real-time. Facilitated group discussion will allow the focus group to review their collective responses and permit further elaboration upon their individual and collective feedback.

- **Task 3.2: Unstructured Exploration.** (RQ3)

During unstructured exploration, the research team will give the focus group some tips for effective prompts, and then encourage them to explore questions or tasks of their choosing. The chatbot will allow us to capture the prompts for further analysis. As with the structured exploration, individual evaluation and assessment will be collected via real-time, live feedback to record and capture instant impressions. During the group discussion, we will be particularly interested in eliciting focus group feedback on the limitations of chatbot performance, which will inform subsequent model tuning and modifications to the user interface and prompts in Phase 4.

### 4. Phase 4: GPT Performance Analysis & Re-tuning (June-November 2025)

- **Task 4.1: Comparative Analysis of Non-AI Ground Truth** (RQ2)

A comparative analysis of the responses from Non-AI captured as ground truth with the DTA aggregate data analysis and the responses from the GPT-powered conversational chatbot for the questions asked by the focus group from Working Meeting I is performed in the steps below:

- Compile and Consolidate Data:
  - Non-AI Ground Truth Data: Gather the data collected in Phase 2, which serves as the baseline for what traditional, non-AI methods can reveal about the DTA dataset.
- Define Comparative Metrics:
  - Accuracy: Measure how closely the GPT's responses match the established non-AI ground truth, focusing on the correctness of information and aggregated data results.
- Conduct the Comparative Analysis
  - Quantitative Analysis: Use statistical methods to compare the accuracy and efficiency of the GPT's answers against the non-AI ground truth benchmarks.
- Document Findings and Insights
  - Report on Comparisons: Create a detailed report documenting the comparative analysis, highlighting areas where the GPT excels or falls short compared to non-AI methods.

- **Task 4.2: After Working Meeting I GPT Performance Evaluation.** (RQ1, RQ3)
  - Establish Evaluation Criteria: Based on the objectives set before Working Meeting I and the feedback received, define clear criteria for evaluating the chatbot's improved performance. This includes evaluating validity, reliability, and reproducibility.
  - Post-Meeting GPT Performance Data: Assemble all data regarding the GPT's performance after the responses captured from Working Meeting I.
  - Analyze User Feedback: Analyze evaluations collected during Working Meeting I, to identify data and GPT performance gaps..
  - Stakeholder Feedback
    - Present Findings: Share the comparative analysis report with stakeholders, including project team members, focus group participants, and funding bodies.
    - Gather Feedback: Solicit feedback on the findings, particularly regarding the implications for future iterations of the project.
  - Plan for Further Improvements
    - Identify Action Items: Based on the comparative analysis and stakeholder feedback, list specific areas for further improvements in both the dataset and the GPT.
    - Outline Next Steps: Develop a plan for addressing these areas in subsequent phases of the project, including potential additional meetings, data enhancement, or model re-tuning.
  - Steps to refine chatbot responses: Based on the evaluation, outline steps for further refining the chatbot's responses. This may involve additional rounds of feedback gathering, dataset updates, or model re-tuning by looking for targeted measures with respect to data annotation, metadata, model features, and model functionality.
- **Task 4.3: Re-prepping of the Dataset.** (RQ1)
  - Analyze Feedback: Identify common themes and specific issues raised by focus group members. Determine which aspects of the dataset need refinement to improve the chatbot's understanding and response accuracy.
  - Update Dataset: Based on the feedback analysis, make necessary additions, deletions, or modifications to the dataset. This could involve adding new records, correcting inaccuracies, or enhancing metadata for better contextual understanding.
  - Document Changes: Keep a detailed record of all changes made to the dataset for transparency and future reference.
- **Task 4.4: Re-tuning the Model.** (RQ3)
  - Review Focus Group Feedback: Focus on the feedback concerning the chatbot's performance, including how well it understood and responded to queries, and any noted biases or inaccuracies. This includes evaluating validity, reliability, and reproducibility.
  - Identify Re-tuning Needs: Based on the feedback, pinpoint specific areas where the model's performance can be improved. This may include adjusting its sensitivity to cultural nuances, improving its understanding of context, or enhancing its ability to generate relevant responses.
  - Implement Adjustments: Apply the necessary modifications to the model's configuration. This could involve re-training the model with the updated dataset, adjusting its parameters to refine its responses, or incorporating new rules to reduce bias.
  - Test Adjustments: Conduct thorough testing to ensure that the re-tuning effectively addresses the feedback from Working Meeting I. This testing should mimic real-world usage scenarios as closely as possible.
- **Task 4.5: Documentation and Reporting.** (RQ1, RQ2, RQ3)
  - Throughout Phase 4, it is essential to document every step taken, from the initial feedback analysis through to the final performance evaluation. This documentation should include:
    - Detailed notes on feedback themes and how they were addressed.
    - A log of changes made to the dataset and the model, including rationales for each change.
    - Results from the performance evaluation, highlighting improvements and ongoing challenges.
    - Recommendations for future adjustments and areas for further research or development.
- **Task 4.6: Disseminate Interim Activities and Findings.** (RQ1, RQ2, RQ3)

After the first working meeting, there will be many opportunities to disseminate information about project activities and findings, to both formal and informal audiences. We anticipate that the partnership with Kennard and the Maryland State Archives will allow us to prototype community learning and engagement, which could be replicated by other local and regional library and cultural heritage organizations.



- Prepare panel presentations on interim activities for IEEE Big Data/Computational Archival Science in December 2024. This dissemination opportunity occurs before Phase 4. It will allow us to introduce the project to a global research community, which can help drive regional and local interest in the dissemination of products and channels in Phase 4 and Phase 6.
- Findings from the first working meeting will be disseminated directly to the Kennard African American Cultural Heritage Center, the Chesapeake Heartland project, and the Maryland State Archives for incorporation into their ongoing research, displays, and community-based cultural heritage activities
- Develop content for real-time project updates and community learning to share via online media of project partners (see also Task 6.4)
- Maintain public GitHub repository of survey methods, operationalizing qualitative data, annotation, metadata creation and enhancement, and model turning.
- Prepare paper(s) and/or panel presentations for the Society of American Archivists annual conference in August 2025
- Prepare papers(s) and/or panel presentations for IEEE Big Data/Computational Archival Science in December 2025

## 5. Phase 5: Focus Group Working Meeting II (November 2025 - January 2026)

We will reconvene the focus group in person at Kennard for a second working meeting in January 2026. The research team will take the focus group once again through the steps of structured and unstructured exploration, as described in Phase 3, and elicit real-time evaluation of usability, contextual accuracy, bias mitigation, and use case relevance. This will permit comparison with prior results. The group discussions at this Working Meeting will focus on identifying technical and governance needs for implementing this and similar tools in various use cases, as well as defining areas for future research. (RQ2, RQ3)

## 6. Phase 6: Wrap-up (February-August 2026)

- **Task 6.1: Consolidate Research Findings and Insights.** (RQ1, RQ2, RQ3)
  - Synthesize Data: Compile and analyze data from both working meetings, the dataset re-preparation, and model re-tuning activities.
  - Document Lessons Learned: Highlight key lessons regarding the GPT's usability, contextual accuracy, bias mitigation, and relevance to various use cases.
- **Task 6.2: Evaluate GPT Model Performance and Project Goals.** (RQ2, RQ3)
  - Compare Pre and Post-Tuning Performance: Evaluate the GPT's performance before and after the second tuning, using metrics established in earlier phases.
  - Assess Goal Achievement: Measure the project's success in meeting its initial goals, especially those related to improving access to and use of library and archives collections.
  - Identify Remaining Gaps: Document any areas where the project fell short, including technical limitations, unmet user needs, or unresolved issues related to bias and contextual accuracy.
- **Task 6.3: Define Areas for Future Research.** (RQ1, RQ2, RQ3)
  - Highlight Technical Innovations: Based on the project's findings, outline potential areas for technological innovation or further development in the use of AI for library and archival collections.
  - Propose Governance Frameworks: Suggest frameworks for the ethical and effective governance of AI tools in library and archival settings, considering privacy, data protection, and bias mitigation.
  - Recommend Future Research Topics: Identify specific topics for future research, inspired by the gaps and opportunities discovered during the project.
- **Task 6.4: Disseminate Final Findings.** (RQ1, RQ2, RQ3)
  - Prepare Final Report: Compile a comprehensive final report that includes methodologies, findings, lessons learned, and recommendations for future work.
  - Final Dissemination Activities: Plan and execute a series of concluding dissemination activities, such as webinars, conference presentations, and publications in relevant academic journals or industry newsletters, to share the project's outcomes with a broader audience.
  - Prepare paper(s) and/or panel presentations for the Society of American Archivists annual conference in August 2026.

Our project, by its very design, lends itself to live and iterative dissemination throughout. The outcomes of our Working Meetings will immediately be disseminated to the Kennard Center and Chesapeake Heartland projects for incorporation into displays, and community-based cultural heritage activities (Task 4.6). In addition to traditional dissemination activities mentioned in Task 4.6 and Task 6.4, we will engage in community-centric opportunities in real-time via social

media channels of the project partners. This will allow us to document project updates and related activities, engage community learning, and promote replication by local cultural institutions in the US and abroad. Content will include podcasts, interviews with descendants related to the ads, and learning modules for educators. This is illustrated through a recent Feb. 2024 example of [community engagement and dissemination](#) from Kennard African American Cultural Heritage Center.

### 3. Diversity Plan

We are committed to fostering diversity and inclusivity in our project, directly and indirectly. To achieve this, we will collaborate closely with **Christopher E. 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. Mr. Haley will support the exploration and creation of an AI-ready DTA Collection.

In addition, we will collaborate with **Clayton Washington**, President of the Kennard Alumni Association. This non-profit runs the [Kennard African American Cultural Heritage Center](#) in Centreville, Queen Anne's County, Maryland. The KAACHC was created and opened 20 years ago, with a small group of people starting a grassroots initiative to revive and restore the original Kennard High School and create a community center. This center has become the hub for African American History programs and events while showcasing the exhibits of its African American History Museum and providing educational and cultural learning opportunities. The center also provides an additional community meeting place and is available for event and meeting rentals, which we intend to use for the Working Meetings in Phases 3 and 5. Mr. Washington is also a member of the board of the [Chesapeake Heartland](#) project out of Washington University in Chestertown, Queen's County, Maryland, also on the Eastern Shore, whose mission is to preserve, digitize, interpret, and make accessible materials related to African American history and culture. The DTA and other LoS collections include many records associated with enslaved and formerly enslaved people in Queen Anne's County. We will host two Focus Group Working Meetings at Kennard and bring together expert users to test the generative AI technology. These would be residents who have backgrounds in working with history, preservation, and education on the Eastern Shore.

The lessons learned from this project apply to archival or historical collections representative of other localized and historically marginalized communities. Important cultural histories often lie within these collections, and otherwise would remain undiscovered. Our applied research process of engaging a minoritized community can be replicated by and/or with other minoritized groups to explore and uncover their contribution to the historical record.

Our research team is interdisciplinary and represents diversity in experience, heritage, and gender. Our resumes summarize the mix of research and professional experience that we bring to the project. Team members have heritage on four continents and represent three ethnic groups. We are male and female.

### 4. Project Results

The MSA's DTA is a rare and unique collection documenting the buying and selling of enslaved people in Maryland. 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 to be reused by other institutions. We will promote these products through our ongoing and real-time project dissemination activities, including listservs, and the AI-Collaboratory website.

Our partnership with the Kennard Center enables immediate incorporation of project outcomes into additional local research activities (via the Chesapeake Heartland project) and cultural heritage displays and activities. The dissemination and prototyping done by the Kennard Center can provide the basis for training other cultural heritage professionals, educators, and community-based groups on using GPT models for formal and informal education and outreach.

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 GPTs to libraries and archives, where we would incorporate lessons learned from this project. Initially piloted at the U. Maryland iSchool, the course will be disseminated through the IMLS-funded TALENT Network ([RE-252287-OLS-22](#)) and its network of iSchools and HBCU partners and its CASES platform (Computational Archival Science Educational System at <https://cases.umd.edu/>). Our goal continues to be to create a durable, diverse, and multidisciplinary national community focused on developing archival and library educators and practitioners who can be future digital leaders.



## Digital Products Plan

This digital products plan aligns with the IMLS requirements for digital products, focusing on the types of digital products to be created, their availability, access rights, and sustainability. This plan integrates the project's specific details, such as selecting GPT models, data preparation, model training, focus group working meetings, performance analysis, and dissemination, ensuring that the plan reflects the project's comprehensive approach to utilizing generative AI in library and archival settings.

### 1. Type of Digital Products:

This plan outlines the types of digital products to be developed in the "Harnessing Generative AI to Support Exploration and Discovery in Library and Archival Collections" project by the University of Maryland iSchool, in collaboration with Maryland State Archives and the Kennard African American Cultural Heritage Center. The project leverages generative AI technologies to enhance the accessibility and utility of the Domestic Traffic Ads (DTA) collection, ensuring innovative, user-centered archival exploration and discovery. Below is a list of the specific digital products. Please refer to the **“Data Management Plan”** document submitted with this application for additional information. We wish to emphasize that many of these products will be created throughout the project, not just at the end. Tasks 4.5 “Documentation and Reporting” and Task 4.6 “Disseminate Interim Activities and Finding” in particular, express the incremental nature of our digital product creation approach.

- Phase 1: GPT Selection & Focus Group Input for Model Design
  - Conversational chatbot
    - A digital code repository comprising a full-stack web application, including the user interface for the chatbot
  - Focus group design survey
    - A digital survey/questionnaire
  - Paper for Dec. 2024 IEEE Big Data / Computational Archival Sciences (CAS) workshop
- Phase 2: Data Preparation & Model Training
  - Full DTA Collection augmented with additional fields, curated for AI use, and incorporating user input
    - Optical Character Recognized (OCR) text from High-resolution images of historical advertisements will be accompanied by extensive metadata detailing each ad's historical and cultural context. Formats include CSV and TXT files for OCR text and metadata, adhering to interoperability standards. A dataset will be created and enriched with metadata reflecting community input and tailored for AI interaction. This dataset will be the foundation for the AI-driven chatbot, equipped to provide informed, contextually accurate responses to user queries.
  - Non-AI ground truth of DTA aggregated data
    - A collection of traditional data analysis outputs for validating AI-generated insights, including statistical analyses, visualizations, and narrative interpretations, provided in standard formats like CSV or PDF. A comprehensive analysis report comparing AI-generated insights against non-AI ground truth data will be produced, highlighting the AI's accuracy, efficiency, and areas for improvement. This would also include the results comparing the GPT model outputs to the ground truth outputs. This document will inform ongoing refinements to the AI model, enhancing its reliability and user relevance.
  - GPT for DTA conversational chatbot
    - A detailed evaluation of the GPT models will be created, emphasizing those with robust privacy controls and sensitivity to the historical context. This report will guide the AI framework's development, ensuring it aligns with the project's ethical and operational standards.
- Phase 3: Focus Group Working Meeting I
  - Focus group working meeting I assessment data and feedback
    - Structured datasets capturing user interactions and feedback on the AI tool, stored in CSV format,

with comprehensive documentation outlining the data collection methodology.

- Paper for Society of American Archivists (SAA) 2025
- Phase 4: GPT Performance Analysis & Re-tuning
  - Comparative analysis of non-AI ground truth
    - Comparative analysis datasets of different GPT models, including performance metrics, evaluation criteria, and selection justifications, presented appropriately for interoperability.
  - Revised dataset
  - Revised GPT model
  - Documentation
    - Comprehensive guides and metadata descriptions ensuring clarity on data use, structure, and collection processes, provided in CSV or PDF format.
  - Paper for Dec. 2025 IEEE Big Data / CAS workshop
- Phase 5: Focus Group Working Meeting II
  - Paper for SAA 2026
- Phase 6: Wrap-up
  - Final report
    - Detailing the project's outcomes, methodologies, lessons learned, and recommendations for future AI applications in archival settings will be produced. This report will serve as a valuable resource for stakeholders interested in the intersection of AI, libraries, and archival science.
- Products after Project Wrap-up: After the project is deemed complete, there would be digital products created that comprise of webinars, presentations, academic papers, exhibits produced to disseminate the project results and findings to the public.

## 2. Availability & Standards

The digital products produced except for the DTA advertisement dataset, focus group user feedback, will be freely accessible, adhering to open-access principles and IMLS's commitment to a broad public utility, such as accessible to a wide range of users, including researchers, educators, students, and the general public. The project will ensure the long-term availability and preservation of these digital outputs with clear documentation, open formats, and sustainable hosting and maintenance strategies, as mentioned in the Data Management Plan. We will commit to open standards and interoperable formats, enabling seamless integration with the university's existing digital library infrastructures and ensuring long-term availability and usability of the digital products.

## 3. Access

The digital outputs will be licensed under Creative Commons, ensuring they can be freely used, shared, and adapted while providing clear guidelines on attribution and non-commercial use. This approach respects intellectual property rights and encourages the educational and scholarly application of the project's products. Clear documentation of privacy considerations or cultural sensitivities will be provided with strategies outlined for respectful and ethical handling of sensitive data. We will commit to IMLS's policy on open access, providing the public with free, easily accessible, and reusable digital products. The project's source code will be hosted on GitHub, a publicly accessible website. The datasets and supporting documentation will be hosted on a publicly accessible website and shared storage as part of the AI-Collaboratory network, potentially under the URL - <https://ai-collaboratory.net/projects/heritage-ai>.

## 4. Sustainability

We will implement strategies for the enduring availability of digital products, including regular updates, compatibility checks, and adherence to digital preservation best practices. We will establish a feedback loop with the user community to continually refine and enhance the digital offerings, ensuring they remain relevant, useful, and aligned with user needs and technological advancements. We will create detailed documentation for all digital products, ensuring future users can understand and utilize the data effectively.



## Data Management Plan

### Overview

This Data Management Plan (DMP) outlines the strategies for managing the data generated and utilized during the "Harnessing Generative AI to Support Exploration and Discovery in Library and Archival Collections" project. It aligns with the IMLS requirements, ensuring data is accessible, shareable, and preserved following best practices. The DMP encompasses data collection, documentation, storage, preservation, sharing, and access, specifically tailored to the project's phases and tasks.

### Data Collection

The project will collect rich data encompassing the full Domestic Traffic Ads (DTA) dataset collection, enriched metadata, user interactions, AI-generated responses, and associated metadata. This data includes:

- Historical Ads and Metadata: Digitized ads and comprehensive metadata detailing each item's content, context, and historical significance.
- User Feedback: Critical insights from user interactions with the chatbot, evaluating the usability, contextual accuracy, cultural bias, and use case relevance of the chatbot.- There will also be written and verbal discussion feedback that will be analyzed to provide valuable context for improving the AI model.
- Focus Group Survey: Responses collected from focus group participants before the first working meeting, including valuable data on user expectations, experience, and specific needs. This collection will also incorporate a "CONSENT" form, aligned with Institutional Review Board (IRB) standards, to ensure ethical compliance and informed participation in the initial survey and the subsequent two working meetings.
- Model Selection:
  - Preliminary Research and Shortlisting: Data from an exhaustive review of the latest GPT models, focusing on those with robust privacy controls suitable for sensitive data, including a shortlist of potential models like Llama2, Microsoft Azure's GPTs, AWS's Anthropic Claude, and Google's Gemini.
  - Evaluation Criteria Definition: Information on the criteria set for model evaluation, emphasizing sensitivity to historical context, privacy, security, and customizability, ensuring the model's alignment with the project's ethical and functional standards.
  - Final Model Selection: Data related to the selection of the optimal model, which will be determined based on thorough testing to ensure it meets the project's specific requirements in terms of data sensitivity, customization, and privacy standards.
- Generate Non-AI Ground Truth of DTA Aggregated Data: We choose to work with Jupyter Notebooks (there are many options though). Data will be gathered using Jupyter Notebooks to conduct traditional, non-AI data science approaches, including statistical analyses and Python code executions, to process and analyze the dataset thoroughly. Data will be collected by creating visual representations like charts, graphs, and maps, alongside documenting insights and observations that detail the significance of the findings. These data would be assembled as a comprehensive ground truth dataset for baseline comparisons, structured to juxtapose directly against GPT-generated outputs, summarizing key findings and metrics. Data associated with the peer review process would also be collected and stored to ensure the accuracy, reliability, and validation of the traditional data analysis results.
- IRB Compliance and Consent: As this project will be submitted to the IRB for their review, detailed attention to ethical standards, with a robust consent process will be integrated into the workshop survey, ensuring that all data collection methodologies are transparent and approved, safeguarding

participant privacy, and adhering to research integrity. We expect to ask for an IRB exemption.

### **Roles and Responsibilities**

The Advanced Information Collaboratory team will oversee data management, with specific members designated for data curation, security, and compliance monitoring. Responsibilities include ensuring data integrity, implementing privacy measures, and facilitating data sharing and preservation.

### **Data Storage and Security**

Data will be stored on secure university-owned cloud servers with data encryption at rest and during transit, encrypted backups, and protection against unauthorized access, data loss, and breaches.

### **Data Preservation and Sharing**

Upon project completion, the dataset will be preserved in the university's digital repository, guaranteeing long-term preservation. This would be a repository created under the overarching project - <https://ai-collaboratory.net/projects/heritage-ai>. Metadata standards will be employed to enhance the dataset's findability and accessibility. The data will be shared under an open license, allowing for research reuse and replication, subject to legal or ethical restrictions. The original DTA digitized dataset is available from the Maryland Legacy of Slavery database website [here](#) and this project will only host the cleaned and updated data with additional metadata used for this project specifically.

### **Access and Data Sharing**

Consistent with IMLS's open-access policy, the input data, model documentation, and aggregate user data will be made publicly available through the university's data repository, ensuring broad access for research and educational purposes. The project's source code will be hosted on GitHub, a publicly accessible website. The cleaned DTA with metadata, quantitative user feedback, and surveys, aggregated qualitative user feedback and surveys, comparative analysis datasets, and supporting documentation such as reports will be hosted on a publicly accessible website and shared storage as part of the AI-collaboratory network, potentially under the URL - <https://ai-collaboratory.net/projects/heritage-ai>. Disaggregated qualitative data will only be made available with consent and upon redaction of private, confidential, or culturally sensitive information.

### **Data Formats and Documentation**

Data will be curated in interoperable formats, ensuring usability across various platforms. Comprehensive documentation will accompany the dataset, detailing the data collection, structure, and coding schemes used, facilitating reproducibility and secondary data use.

### **Ethics and Legal Compliance**

The project commits to upholding ethical standards and legal compliance, particularly in handling sensitive historical data. It includes mechanisms for addressing copyright, privacy, and data protection, ensuring the project's alignment with applicable laws and ethical norms.

### **Data Management Review and Adaptation**

The DMP will be periodically reviewed and updated to adapt to project evolutions, technological advancements, or changes in best practices and regulations. This iterative approach ensures the DMP remains relevant and effective throughout the project lifecycle. A timeline will be established for regular reviews of the DMP, including pre- and post-community Focus Group Working Meeting evaluations, and at the end of each project phase. Mechanisms for monitoring adherence to the DMP, including stakeholder engagement and feedback loops, will be provided in detail.