School of Information, University of Michigan

Improving Metadata Quality and Minimizing Disclosure Risk with Human-AI Data Curation Pipelines

We propose an applied research study that addresses the potential of generative artificial intelligence (GenAI) to augment manual curation in data archives. Our project addresses objectives 3.2, listed under “Goal 3: Improve the ability of libraries and archives to provide broad access to and use of information and collections.” We request $749,965 from IMLS. Our proposal sits at the intersection of artificial intelligence and digital curation. We will generate insights about how GenAI may (a) generate metadata to improve research dataset discovery and evaluation and (b) reduce disclosure risks for individuals and communities represented in research data. We will share our findings, code, and documentation broadly with research and practice communities.

Project Justification

In our recent IMLS-funded study on curatorial actions in digital collections, we identified two significant challenges archives face in managing digital collections. First, detailed descriptive metadata is necessary for facilitating findability and reuse, but creating that metadata requires time and expertise that are scarce and expensive. Second, the most time-consuming curatorial actions in a data archive are related to disclosure risk review. To address these two challenges, we address the following research questions: (1) Metadata drafting and evaluation: how can GenAI tools facilitate data producers and curators in drafting metadata? and (2) Disclosure risk review: how can computational tools help data producers and curators identify and handle potentially sensitive direct and indirect identifiers in data?

Metadata creation is a key component of ensuring data are discoverable and re-usable. High-quality metadata increases the chances of users receiving ideal dataset recommendations since most dataset search systems are based on metadata; it is difficult to avoid the problem of ‘garbage in, garbage out’ without high-quality metadata. Data reuse is likely to increase with better discoverability, fostering open science. Improved data discoverability also benefits data depositors as they can get more citations from data reuse (Piwowar & Vision, 2013). Further, Metadata management remains a pressing concern for data repositories as improving metadata allows them to gain more trust from their users and achieve higher user satisfaction, thereby leading to more data reuse. Among various factors (e.g., data completeness, accessibility, ease of use, data credibility) that affect data reusers’ satisfaction (Faniel et al., 2016), metadata quality is most significant for data curators because they can actively intervene to control it.

Data reusers feel more satisfied when provided with high-quality metadata because they do not need to search for additional resources, which saves time and effort. While perceived concerns about data negatively affect scientists’ intentions to reuse data (Kim & Yoon, 2017), high-quality documentation and metadata relating to legal information and data collection methods may alleviate some legal, methodological, and ethical concerns, thus increasing data trustworthiness and reusability. For instance, users preferred rigorous and clean documentation when evaluating the trustworthiness of repositories (A. Yoon, 2014); they were often looking for specific information about the methodology of the data collection and choices investigators made while collecting the data (A. Yoon, 2014, 2017).

However, generating extensive metadata takes time and expertise from data producers and data curators. Furthermore, manually drafted metadata will inevitably vary in quality because human experts have varying standards and opinions about what metadata should contain. Metadata drafting can be challenging in this situation, as it is difficult to guarantee minimum quality requirements. GenAI has the potential to augment the manual labor of creating metadata while maintaining quality and free data producers and curators to focus on describing the potential uses of the data and data’s particular caveats. We will create human-AI workflows for metadata creation and evaluation. As some data repositories require data depositors to fill in metadata, our proposed tool will be also useful for those without data curation skills.

Our project will experiment with large language models (LLMs) trained and fine-tuned on different texts to characterize their abilities to draft metadata for different datasets. LLMs are pre-trained language

models that use deep learning techniques to process and comprehend natural language (Shen et al., 2023; Zhao et al., 2023). LLMs are trained and fine-tuned on vast amounts of text data, which allows them to learn patterns in unstructured sequences and build a knowledge base of language (Brown et al., 2020; Radford et al., 2019). LLMs offer outstanding advantages over conventional NLP models. In contrast to the conventional approach for NLP tasks, which involves fine-tuning models through supervised learning on small, task-specific datasets, LLMs can effectively perform a wide range of tasks with only a few prompts (Manning, 2022). By providing them with human language descriptions or several examples of the desired task, they can execute tasks for which they were not explicitly trained (Manning, 2022). Both open and closed-source LLMs are widely used for text generation and text-based reasoning. We provide examples that we plan to experiment with in Table 1.

Table 1. LLM Examples

Name	**Creator / Cloud Platform Access Providers** (if have)	**Open Source Status**
Claude	Anthropic / AWS	Closed Source
GPT	OpenAI / Azure	Closed Source; UM has data protection agreement in place
PaLM	VertexAI / Google Cloud	Closed Source
Gemini	VertexAI / Google Cloud	Closed Source
LLaMa	Meta	Open Source
Falcon	Technology Innovation Institute (TII), UAE	Open Source

Data archives face challenges balancing privacy for individuals represented in data and analytic utility for data users. Some approaches to addressing the privacy-utility trade off include introducing noise, generating synthetic data, or employing differential privacy (DP) techniques. Many of the datasets that social scientists employ are relatively small and/or stand alone; there aren't enough observations to preserve utility when introducing noise or generating synthetic data. Too much noise reduces the analytic utility, and too little fails to address privacy risks. Most DP approaches are designed to address a specific set of predefined research questions (Dwork et al., 2006), and when multiple researchers use the data for various questions, fairly allocating privacy budget and the associated DP techniques are challenging (Pujol et al., 2020).

ICPSR offers a different approach – reviewing disclosure risk on data or analysis egress in addition to reviewing the data itself. We refer to processes of reviewing datasets for sensitive data and planning approaches to mitigate their associated risks as "disclosure risk review" (DRR). ICPSR classifies potentially disclosive data as "restricted" and requires potentially reusers to apply for access. Once granted, for especially sensitive data, users must analyze the data within a virtual data enclave that prevents them from downloading or sharing data without manual review of their results. During this manual step, ICPSR staff review all output files for potential privacy risks. These manual reviews are time intensive, and some level of automated assistance could reduce curator effort and catch potential errors. For more details about ICPSR's handling of sensitive data and data about those who apply for access, see our earlier work on privacy impact assessments (Mhaidli et al., 2022).

We propose to investigate how computational tools, including GenAI, can assist data producers and curators with DRR. Our prior work demonstrates that disclosure risk review adds value to datasets, in part by involving curators and archive staff in data review and the many judgment calls required to protect research participants (Thomer et al., 2022).

Project Work Plan

Our project requires two parallel tracks of research. First, in metadata drafting and evaluation, we examine how GenAI can decrease the time it takes to draft metadata and improve the fit between metadata and user queries to boost data discoverability. We characterized curator workflows and data searchers' behaviors in our previous work (Thomer et al., 2022) and used those workflows to identify tasks that GenAI could facilitate. We focus on drafting data description and summarizing descriptive statistics from datasets. For each experiment in both tracks, we will evaluate results with data curators in a lab-based user study. We will also compare the text descriptions and summary statistics with descriptions generated by curators and researchers. We include expert curators and curators-in-training in our project team to ensure that we have the expertise available to evaluate the LLMs' performance on all tasks. We propose two sets of experiments to characterize GenAI's abilities to augment these tasks. We summarize our research questions and indicate which data will address them in Table 2. We provide details about how we'll investigate and evaluate our findings in each section below.

Table 2. Research Questions and the Data Required to Address Them

Research Question	Where Addressed	Data Required
Can GenAI reduce the time required to generate useful dataset summaries?	Experiment Set 1	Curator time [from (Lafia et al., 2021)] GenAI response time Curator and data librarian feedback
Do dynamic variable descriptions improve data discovery?	Experiment Set 1	Curator and data librarian feedback User feedback
Can GenAI produce summary statistics that improve data discovery?	Experiment Set 1	User feedback
Can we automatically detect personally identifiable information in datasets?	Experiment Set 2	Curator and data librarian feedback
Can GenAI assess potential disclosure risk in datasets and analysis output?	Experiment Set 2	Curator and data librarian feedback User feedback

Our studies require engagement with data librarians, curators, and data reusers. We will work with our data impact librarian to develop appropriate experiment protocols for the user studies and then work with the UM IRB office to receive approval. PI Hemphill has a long history of conducting qualitative research and working with UM IRB.

Our goals are to identify uses of GenAI that can improve dataset curation by augmenting, not replacing, curator effort and existing curation practice. We expect our results, especially the models for PII detection and risk assessment, to be useful to data librarians and curators who work with investigators to prepare research data for reuse. ICPSR's curators and disclosure review staff are interested in ways to make their roles more efficient and replicable, and our approach helps with both aspects of their work.

Experiment Set 1: Drafting metadata

Data repositories are challenged by incomplete, short, and inaccurate metadata (see Figure 1) due to staffing shortages, time constraints, and limited technical support (Moulaison Sandy & Dykas, 2016). Also, some data repositories expect data depositors to fill in metadata information, causing metadata quality inconsistencies; some researchers struggle to deposit their data in a repository due to a lack of technical skills and knowledge of metadata creation (Perrier et al., 2020). To address this issue, we introduce a task of description rewriting that uses GenAI to augment low-quality dataset descriptions.

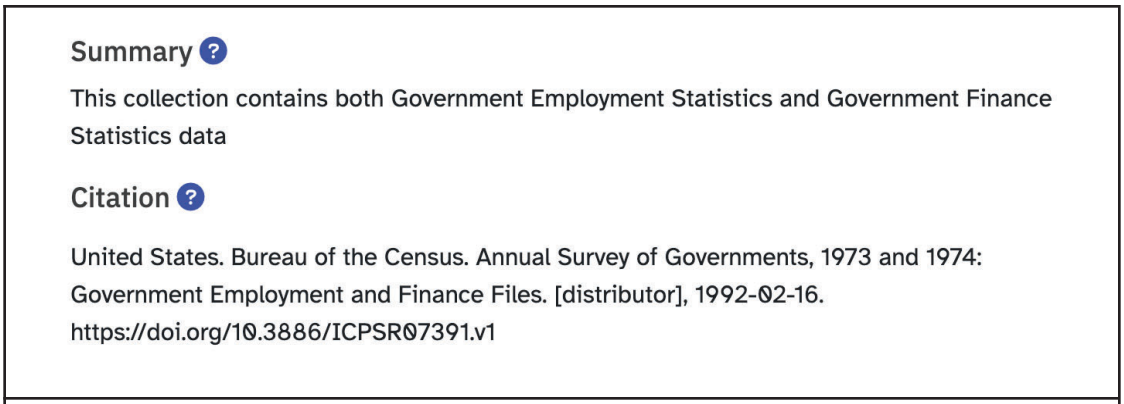

Figure 1. Example ICPSR dataset description. The description (summary) does not elaborate on important contextual information of the dataset.

In these experiments, we study the ways GenAI can augment human efforts to create complete, helpful metadata. First, we will compare different LLM models' performance on a summary drafting task. We piloted this protocol with GPT-4 using non-sensitive data from a survey we conducted. Given the data's codebook and a prompt to summarize the variables, GPT-4 was able to generate a reasonable description. However, its description used too many adverbs (e.g., richly) and confused metadata variables (e.g. start_time) with content variables (e.g., answers to survey questions). We will compare the task performance of several commercial (e.g., GPT-4, PaLM 2) and open-source models (e.g., Falcon, Llama 2). We will also experiment with different prompts to determine whether it's possible to teach an LLM to identify differences between metadata and data and to use more straightforward, academic language.

Our proposed experiments will use non-restricted datasets provided by the Inter-university Consortium for Political and Social Research (ICPSR), which is one of the largest social science data archives. We use only non-restricted data because ICPSR has reviewed them to ensure that disclosure risk is low, and we are unlikely to expose research participants' personal information to proprietary GenAI models. Metadata elements in each dataset include title, description, keywords, funding agency, geographical coverage, and release date. The average dataset description length is 208.84 words (SD = 138.67) when split by whitespace; the maximum and minimum lengths are 2,681 and 0, respectively.

The procedure for the experiment consists of the following stages:

(1) Extract metadata elements from a dataset
(2) Collaborate with GenAI to draft a dataset description based on the given metadata elements
(3) Assess the performance of human-GenAI collaborative writing using automatic and human evaluation metrics

Retrieving metadata elements

This step is common to all experiments in set 1. They all depend on existing metadata for comparison and/or training. ICPSR datasets typically contain a variety of high-quality metadata elements labeled by

human experts. However, some of them lack detailed metadata because data providers did not supply such information in the first place or simply because they were collected a long time ago. In such a case, we will use dataset-related documents, such as codebooks, reports, and publications, to automatically extract metadata information with human-conducted verification. When we piloted metadata extraction with GPT-4 using a codebook, GPT-4 returned fair-quality metadata information (see more details in the Appendix). Although our proposed metadata extraction process requires human operators to confirm the faithfulness of metadata elements captured by GenAI, this human-in-the-loop approach will enable us to save considerable amounts of time and effort constructing a dataset for this experiment.

Drafting Dataset Summaries

In the first experiment, we will ask human curators to draft dataset summaries with GenAI. We will also use various GenAI models to generate summaries for each dataset. We will display the summaries in a web application where curators can compare the summaries, accept and modify either of them, or get new suggestions from GenAI. We will capture the summaries, edits, and time spent on writing and rewriting tasks. We will also analyze how human curators interact with GenAI via our user interface (e.g., consulting GenAI only once versus multiple times, using AI suggestions directly versus selecting the useful portion from GenAI suggestions).

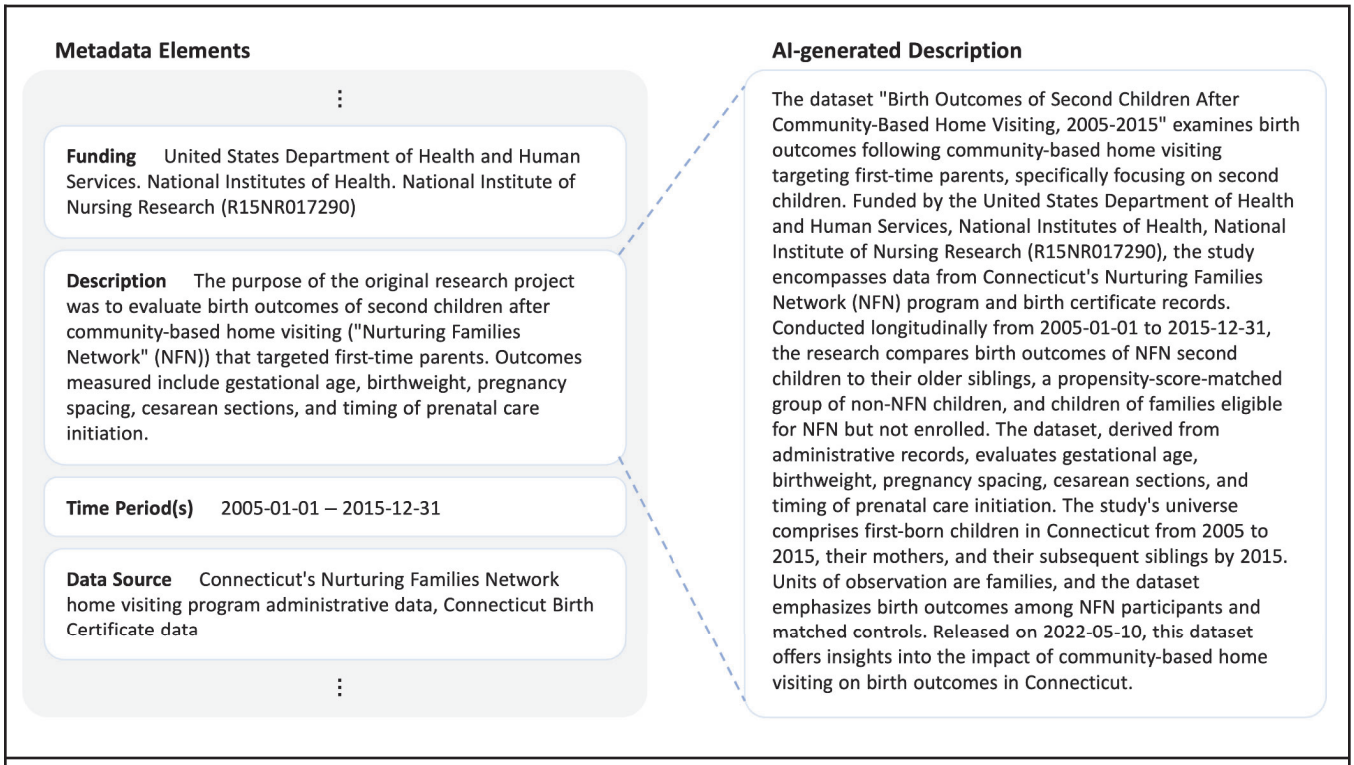

Figure 2. Example AI-generated dataset description (right). GenAI rewrites an original description (left) in given metadata elements. We will provide AI with various metadata elements, such as title, description, funding, geographic coverage, study purpose, time period(s), time method, universe, unit(s) of observation, data source, data type(s), description of variables, and release date.

Evaluating human-AI collaborative outcomes

We will measure how the proposed description rewriting tool assists human curators using both quantitative and qualitative metrics. We will also qualitatively analyze rewritten descriptions by comparing them with the original versions.

First, we will use the following quantitative metrics to measure the quality of AI-generated outputs.

- Edit rate: the number of changes between the original and rewritten descriptions divided by the length of the original description.
- Coverage: the number of metadata elements given to GenAI that are mentioned in the rewritten description.
- Faithfulness: whether an AI-generated description describes every given metadata element correctly without hallucinations.

Further, we will conduct a lab-based user study with human curators and ask them to evaluate their experiences of collaborating with GenAI based on the following criteria:

- Fluency: whether an AI-written suggestion is clear, human-like, coherent, and grammatically correct.
- Future intention to use: our proposed tool's perceived usefulness and potential for future use when drafting dataset descriptions.
- Perceived challenge: how much rewriting a description with generative AI was challenging compared to drafting a description manually.

Drafting Dynamic Variable Descriptions

We also propose dynamically generating variable descriptions for data searchers. While data repositories provide a list of variables of a dataset, users might be overwhelmed given hundreds of variables at once. Furthermore, not all variables in a dataset are equally informative for searchers with different interests and expertise. It is time-consuming for them to manually inspect variables the dataset contains and select those that are relevant to their information needs; often this information is buried in PDF codebooks that require downloading and searching. GenAI has the potential to provide personalized curation service that dynamically describes variables in natural language that mirrors users' search queries.

To generate personalized variable descriptions, we will provide GenAI with different weights for variables in a dataset based on a user's query. Variables relevant to the search query will be highly weighted. For example, assume we have a dataset containing politics- and religion-related variables; if a user types in a religion-related query, AI generates a more detailed description of religion-related variables in the dataset. If a user types in a politics-related query, AI gives more detail on politics-related variables in an updated description. The proposed experiment consists of the following stages:

1. Select potentially relevant datasets for a user's query.
2. Assign weights to variables based on the search query for each dataset.
3. Generate a personalized dataset description for each dataset.
4. Evaluate the performance of AI-generated descriptions using automatic and human evaluation metrics.

Selecting potentially relevant datasets for a user's query

It is computationally expensive to let GenAI rewrite the descriptions of all datasets every time a user types in a new query. We will use a bi-encoder model (msmarco-distilbert-dot-v5) from Sentence Transformers (Reimers & Gurevych, 2019) to select the top 25 potentially relevant datasets for a user's query. This step allows GenAI to generate only 25 descriptions per query.

Assigning weights to variables based on the search query

While ICPSR datasets normally contain hundreds of variables, not all are equally important to data searchers. This step assigns a different weight to each variable in a dataset based on a user's query. When generating a description, GenAI will provide more details about highly weighted variables. For example, in Figure 3, family- and bullying-related variables are highly weighted according to a user query so GenAI will describe more about them if a dataset contains such variables.

Generating a personalized variable description using AI

In this step, we will ask GenAI to rewrite the description of each dataset in a set of 25 datasets chosen in the first step using the calculated weights. Figure 3 illustrates how AI incorporates variable-related information reflecting a user's query into a rewritten dataset description.

Evaluating the quality of AI-generated variable descriptions

We will perform an automatic evaluation that tests whether our proposed tool successfully extracts key variables reflecting a user's search query from a given dataset. To do so, we will create a test collection containing ICPSR dataset-paper pairs as ground truth data. Each ICPSR dataset in the test collection is paired with one or more research papers that analyze it. We will use research questions (or titles) from papers as search queries, and our prototype tool will generate a description of variables given that research question. We will compare variables mentioned in the AI-generated description and those used in the paper to measure the coverage of study variables. This approach effectively measures the overlap of two sets of variables, one from an AI-generated description and the other from the published paper that addressed a specific research question with a specific dataset.

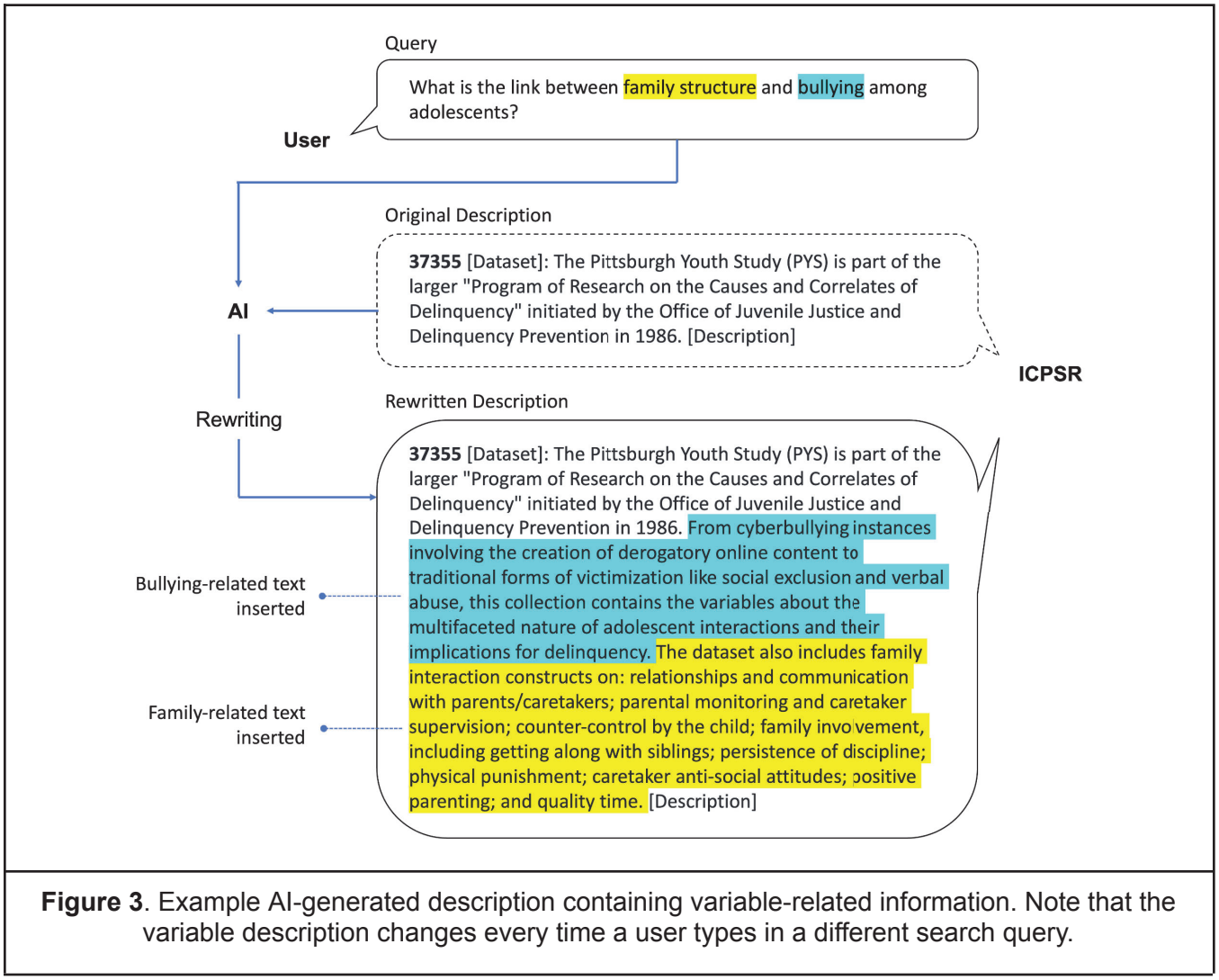

Figure 3. Example AI-generated description containing variable-related information. Note that the variable description changes every time a user types in a different search query.

Drafting Summary Statistics

In another attempt to select and summarize relevant variables, we will ask GenAI to generate summary statistics such as frequency distributions for demographic variables. Data reusers often ask whether a dataset's sample contains sufficient respondents of particular groups (i.e., adequate subsamples of non-white respondents), and descriptive statistics are not a standard component of dataset

documentation. Some researchers include frequencies of each response in their codebook, and datasets with this more complete documentation are more likely to be used. While users can generate descriptive statistics using statistics packages, saving users this step will reduce the time it takes to decide if a dataset is appropriate for their research and minimizes disclosure risk. To calculate descriptives themselves, users must download data, identify the relevant variables, and run their statistics commands. They cannot do so with restricted-use data (i.e., data with sensitive personally identifiable information or that is under limited-use contracts) because they cannot download the data directly or without going through a lengthy application process. GenAI may be able to identify relevant variables, quickly calculate frequencies, and even graph two-way relationships in ways that help researchers quickly understand the data's samples, structure, and fitness-for-purpose without requiring downloads and before they apply for restricted-use data.

Researchers commonly use built-in commands from common statistical packages (e.g., summary() from R (R Core Team, 2022)) to generate summary statistics. However, these tools require direct access to the data, programming knowledge and can be time-consuming, especially for users who are not familiar with these environments. Some datasets also contain hundreds or thousands of variables, and wading through their summaries is not a good use of researcher time. Restricted-use datasets cannot be directly accessed, and so researchers cannot summarize them without applying for access and receiving approval. While our experiment includes only non-restricted data, it's likely that our summary generation techniques will translate to restricted use data. We reserve investigations about generating privacy-protecting summaries for future work or potential inclusion in experiment 2.

GenAI can streamline this data summary process by automatically identifying relevant variables and calculating descriptive statistics. For example, GenAI could generate a summary report that includes frequency distributions, mean and median values, and even graphical representations such as histograms or box plots for key demographic variables. This automation saves time and makes statistical analysis more accessible to researchers without extensive statistical software expertise. Example output from GenAI might include a comprehensive summary report with insights into the dataset's demographic composition, highlighting areas where subsamples may be underrepresented, which could affect the study's validity or applicability.

Experiment Set 2: Enhancing disclosure risk review and suggesting privacy-friendly data use plans

ICPSR uses virtual data enclaves and manual disclosure risk review to mitigate these types of privacy risks for sensitive data. In our second set of experiments, we test ways to use GenAI to improve DRR by reducing the risks data reuse poses to research participants and the time DRR takes to complete.

ICPSR currently relies on a manual review process in which experts meticulously analyze datasets to identify any potential privacy risks, a method that is both time-consuming and resource-intensive. This process includes checking for direct identifiers, assessing the risk of indirect identifiers leading to re-identification, and ensuring that the data complies with legal and ethical standards.

We focus on three types of privacy attacks common for personally identifiable information: identification, inference, and linkage (Liu et al., 2021). Anonymization is a common approach for addressing identification (sometimes called "re-identification") risks, but prior research has shown that anonymization alone is not effective at preserving research participants' identities, even when following standards such as HIPAA Safe Harbor (Sweeney et al., 2017). Synthetic data has shown promise for protecting privacy against all three types of attacks but can be labor- and resource-intensive to generate. Generating differential privacy mechanisms for multiple stakeholders with unknown research questions is a challenge that future work could pursue. We address two specific tasks that are common components of nearly all these privacy protection approaches:

- Personally identifiable information (PII) detection – finding PII in datasets; and
- Risk prediction – assessing the risk of re-identification or exposure of that PII given a set of analysis output.

For object detection in the context of finding PII within datasets, several methodologies could be employed. For instance, techniques such as text mining and pattern recognition can be used to scan datasets for direct PII (e.g., names, social security numbers) and indirect PII (e.g., combinations of birthdate, zip code, and gender). GenAI provides in-context learning techniques through prompt engineering or fine-tuning open source models that improve our ability to recognize and flag potential PII accurately and automatically, thereby streamlining the initial steps of the DRR process. Using GenAI in this way also provides a reusable and documentable process; we can automatically track any PII detected and any data transformations suggested and then use them to evaluate the tool's effectiveness and provide transparency reports about our DRR methods.

Risk prediction involves assessing the likelihood that a given piece of information could lead to the re-identification of an individual. This assessment can be based on the uniqueness of the data points within the dataset and their availability in external datasets, which could be used for linkage attacks. Statistical models and machine learning algorithms can be developed to estimate these risks by analyzing the distribution of the data and comparing it with known external sources.

We will test various machine learning approaches – including entity recognition, GenAI prompts, and fine-tuning – to identify potentially sensitive data and unsuitable (or risky) uses and combinations of datasets. Based on the DRR results, we can make suggestions about ethical and appropriate data use suggestions. Common approaches include data anonymization (Domingo-Ferrer et al., 2022), where direct and indirect identifiers are removed or obscured, and synthetic data creation (Bellovin et al., 2019), where the data is generated to mimic the statistical properties of the original data without including any real individual's information. Deep learning and adversarial neural network-based methods (Abay et al., 2019; J. Yoon et al., 2020), which share similar technical foundations of generativity and word vectorization natures as GenAI, have proved to be especially useful in anonymization and data synthesis, and therefore GenAI-based methods have high potential. We provide two simplified examples of using GPT-4 to identify PII and do a basic risk assessment in the appendix.

To demonstrate GenAI's ability to reason about datasets, we provide two simple examples of PII detection and risk assessment. In the first example, we generated example data about how many hours per day individuals use imaginary internet sites. In the second, we use the first five rows of data from the University of California, Irvine (UCI) COVID-19 Study (Silver et al., 2024). The UCI COVID-19 Study has already been through ICPSR's review processes, and potentially disclosive variables have been manually masked. For both examples, we use a simple prompt:

> Here's a dataset:
>
> [rows of data]
>
> Can you tell me if there's potentially disclosive data in that dataset?

The structure of the responses GPT-4 generated were also similar. First, it proposes definitions for variables, and then it provides an estimate of the disclosive risks of that variable on its own and in combination with other variables. For the synthetic data, we did not provide column headers, and GPT-4 correctly inferred the variable name and data (e.g., first name) and identified them as risky – "The dataset you provided contains data that could compromise an individual's privacy, specifically their full name and email addresses." In the UCI COVID-19 Study, GPT-4 recognized that variables were masked and identified free-text variables that may contain disclosive information. In both cases, it provides recommendations for anonymization and data use practices to protect individuals in the data.

Personnel and Resources

Principal Investigator Libby Hemphill will commit effort each year to the project and will be responsible for mentoring PhD students and project staff, securing appropriate computing resources, and setting the research agenda. She has led projects on data curation (Fan, Lafia, Wofford, et al., 2023; Hemphill et al., 2022; Lafia et al., 2021; Lafia, Thomer, et al., 2023; Mhaidli et al., 2022; Thomer et al., 2022), data recommendation (Fan, Lafia, Li, et al., 2023; Lafia, Million, et al., 2023), and generative AI (Gao et

al., 2023; Li et al., 2023, 2024). She will actively work on the project during the academic year, and her existing academic appointment will cover her salary for that period. The summer support requested here will ensure she can continue working on the project throughout the year. We are requesting support for 2 UMSI PhD student research assistants, and they will be responsible for implementing the research tasks, drafting publications, and generating sharable code and documentation. A master's student research assistant will lead evaluation efforts in all years; this assistant will be a digital curation student in the UMSI master's program. The data impact librarian and curators at ICPSR will also be included in our evaluation activities.

Diversity Plan

In aligning with the principles of our research plan, we are committed to dismantling barriers to data access and analysis and creating a richer, more inclusive body of knowledge in the field of digital curation and data sharing. Our project aims to make data more discoverable and usable and to empower individuals and institutions, regardless of their computational proficiency or resources, to better understand and manage disclosure risks. The efficient and replicable processes we develop for generating metadata and reviewing disclosure risk will make it possible for curators and data librarians to do more with less. Our hope is that the efficiencies GenAI provides free these experts to work more closely with data and users and spend less time and computing energy on drafting documentation. Similarly, improving data discovery likely diversifies the group of users who can reuse data. Our experiments to generate bespoke variable descriptions, for instance, communicate with users in natural language rather than the specialized, often jargon-filled, language of the original dataset documentation. We expect that these natural language descriptions reach more data users.

Our commitment to diversity extends beyond our tools. We aim to understand and address the disclosure risks that affect historically marginalized groups, ensuring that data privacy considerations afford them equal protections. Most disclosure risk mitigation now focuses on individual risks, and we recognize that sometimes the risks of disclosure are for a community. Our disclosure risk experiments will help identify datasets and variable combinations that pose risks for groups in addition to individuals. Community risk is a common concern expressed by data providers when asked to archive their data with the Resource Center for Minority Data at ICPSR. Part of our motivation for conducting these experiments is to find safer ways to facilitate access to data that includes more diverse voices without putting those voices at risk.

Our team includes members of sexual orientation and gender minorities, ability groups, and differs along racial and country of origin demographics. PI Hemphill has over 10 years of experience successfully recruiting and mentoring students from historically marginalized communities and will continue that work with the support of the UMSI Diversity, Equity, and Inclusion office.

Project Results

Our project will produce the following deliverables:

- Peer-reviewed articles that explain our generative AI experiments and their results.
- Well-documented code for using generative AI to draft metadata for non-sensitive datasets.
- Peer-reviewed articles that present the results of our efforts to augment disclosure risk review with artificial intelligence tools.
- Well-documented code for using artificial intelligence to detect potentially disclosive information in datasets.
- Fine-tuned pretrained machine learning models.
- Presentations for researchers and archivists that demonstrate the AI augmentation approaches we evaluate.

Our team has presented at colloquium at archives and graduate schools, to the ICPSR Biennial Meeting, and at professional conferences such as IASSIST and RDAP. We expect to continue this direct engagement with practitioners to receive input and feedback from the data curation community and to spread the word about our tools.

Schedule of Completion

We have divided the work across three academic years. The first two years will follow a similar pattern of experiment planning and testing the fall, data collection in the winter, and review and publication in the spring/summer terms. PI Hemphill will lead the project and will be engaged in each activity. Throughout the first two years, the PhD GSRAs will share responsibility for research activities. We have two current PhD students who are likely to work on these projects – one on dataset summaries, summary statistics, and PII and the other on dynamic variable descriptions and data use recommendations. Depending on their dissertation timelines, the GSRAs may change. The MSI research assistant will be hired during winter or spring/summer terms to manage experiments and assist with outreach and engagement; we expect them to work on experiment recruitment while the PhD GSRAs are finalizing experiment protocols. These will be parallel rather than serial activities. PI Hemphill and the MSI RA will work to complete our data distribution and archiving plans in year 3.

	Year 1			Year 2			Year 3		
	Fall 2024	Winter 2025	Spring Summer 2025	Fall 2025	Winter 2026	Spring Summer 2026	Fall 2026	Winter 2027	Spring Summer 2027
Research Activities									
Drafting Dataset Summaries									
Drafting Dynamic Variable Descriptions									
Drafting Summary Statistics									
Detect PII									
Assess									

Disclosure Risk									
Outreach and Engagement									
Present at IASSIST									
Present at ICPSR Biennial Meeting									
Present at ASIST									
Webinars with ICPSR									
Reporting and Deliverables									
Yearly reports									
Project end report									
Code review and release									
Deposit Data									

Digital Products Plan

Type	Availability	Access	Sustainability
Generative AI prompts *Format*: TXT, CSV **Data analysis code** *Programming languages*: R and Python *Formats*: RMD, PY, IPYNB **Generative AI models** *Programming language*: Python *Formats*: PT, PY, PKL	*Public Websites:* GitHub HuggingFace Zenodo *Delivery strategy:* Code will be made available after it's gone code review by a second member of the project team Publicly available through standard web browsers; manifests will include requirements and dependencies. We will work to minimize dependencies and will document any requirements in our software manifests. Python: Python has well-documented, robust, widely used libraries for training, testing, and saving machine learning models. For instance, we have used pytorch and tensorflow to train and save generative AI models. We expect to continue using those libraries in this project. R: R has well-document, robust, and widely used libraries for analyzing quantitative data. We are	*License:* MIT Permissible licenses, widely used among software developers Our goal is to provide usable software to other archives and research data users; we hope others will build on our software and release under similarly permissive licenses. We will test our models extensively to ensure that they do not expose (or leak) any underlying data. Our other software will not implicate privacy concerns or cultural sensitivities.	*Preservation*: Zenodo ensures the files will be available perpetually. Our requirements files will indicate the software requirements necessary to run the software and will indicate the configuration of the servers on which the code was originally developed. *Maintenance*: We will actively maintain our models as long as we have resources to do so. We hope that by making our models open-source, those that are useful will be maintained by their user communities.

	experts in the tidyverse and expect to use many of its libraries in analyzing our data.		
Research data from experiments *Format*: CSV *Standards*: DDI	DURING THE PROJECT We will store data and models through UM ITS ARC on the Turbo Research storage service. Turbo is a high-capacity, reliable, secure, and fast storage solution. It is tuned for large files (which corresponds roughly to be files in 1 megabyte in size or larger), but it is still capable of efficiently storing small files, such as: word documents, spreadsheets, image files, etc. Turbo enables investigators across the University of Michigan to store and access data needed for their research via a local computer in a lab, office or our High Performance Computing Clusters, such as: Great Lakes, Lighthouse, and Armis2. Additionally, it supports the storing of sensitive data on HIPAA compliant systems, such as the Armis2 cluster managed by ARC. AFTER THE PROJECT Deposited with ICPSR; publicly available	Publicly available, non-restricted license	Deposit with the digital repository of the Inter-university Consortium for Political and Social Research (ICPSR) to ensure that the research community has long-term access to the data. The integrated data management plan proposed leverages capabilities of ICPSR and its trained archival staff. ICPSR will archive the full dataset and its documentation for the long term, supporting the data through changing technologies, new media, and data formats.

Improving Metadata Quality and Minimizing Disclosure Risk with Human-AI Data Curation Pipelines

DATA MANAGEMENT PLAN (prepared with DMPTool)

A.1. Identify the type(s) of data you plan to collect or generate, the purpose or intended use(s) to which you expect them to be put. Describe the method(s) you will use, the proposed scope and scale, and the approximate dates or intervals at which you will collect or generate data.

We will conduct four different experiments and produce performance and evaluation data from each. We expect to generate data during the winter term each year. We expect to release data from our experiments within six months of completing them.

Data from our experiments may include, but are not limited to, metadata generated by humans and generative AI systems, changes made to those generated texts, the time taken to edit and generate texts, similarity scores between texts, and search algorithm rankings.

A.2 Does the proposed data collection or research activity require approval by any internal review panel or institutional review board (IRB)? If so, has the proposed research activity been approved? If not, what is your plan for securing approval?

We will work with the University of Michigan IRB to review our experiment protocols when they are complete. We will not begin interacting with research participants until we have confirmed IRB approval.

A.3 Will you collect any sensitive information?

N/A

A.4 What technical (hardware and/or software) requirements or dependencies would be necessary for understanding retrieving, displaying, processing, or otherwise reusing the data?

We will store our data in plain-text CSV for sharing.

During the project, we will store data on secure Google Drive, UM servers, and in AWS storage services (e.g., S3). Data access will be limited to project team members.

A.5 What documentation (e.g., consent agreements, data documentation, codebooks, metadata, and analytical and procedural information) will you capture or create along with the data? Where will the documentation be stored and in what format(s)? How will you permanently associate and manage the documentation with the data it describes to enable future reuse?

We will include our informed consent documents, experiment protocols, and the resulting data in our ICPSR deposit.

ICPSR will create substantive metadata in compliance with the most relevant standard for the social,behavioral, and economic sciences—the Data Documentation Initiative (DDI). This XML standard provides for the tagging of content, which facilitates preservation and enables flexibility in display. These types of metadata will be produced and archived:

- Study-Level Metadata Record. A summary DDI-based record will be created for inclusion in the searchable ICPSR online catalog. This record will be indexed with terms from the ICPSR Thesaurus to enhance data discovery.
- Data Citation with Digital Object Identifier (DOI). A standard citation will be provided to facilitate attribution. The DOI provides permanent identification for the data and ensures that they will always be found at the URL specified.
- Variable-Level Documentation. ICPSR will tag variable-level information in DDI format for inclusion in ICPSR's Social Science Variables Database (SSVD), which allows users to identify relevant variables and studies of interest.
- Technical Documentation. The variable-level files described above will serve as the foundation for the technical documentation or codebook that ICPSR will prepare and deliver.
- Related Publications. Resources permitting, ICPSR will periodically search for publications based on the data and provide two-way linkages between data and publications.

A.6 What is your plan for managing, disseminating, and preserving data after the completion of the award-funded project?

The research data from this project will be deposited with the digital repository of the Inter-university Consortium for Political and Social Research (ICPSR) to ensure that the research community has long term access to the data. The integrated data management plan proposed leverages capabilities of ICPSR and its trained archival staff.

A.7 Identify where you will deposit the data:

Name and URL of repository: ICPSR http://icpsr.umich.edu

A.8 When and how frequently will you review this data management plan? How will the implementation be monitored?

Yearly. We will review the plan with ICPSR and the project team.