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 (GAI) 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 $749,460 from IMLS. We will generate insights about the intersection of artificial intelligence and digital curation and 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 related to disclosure risk review. To address these two challenges, we address the following research questions (1) Metadata drafting and evaluation: how can GAI 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. However, generating extensive metadata takes time and expertise from data generators and data curators. GAI has the potential to augment the manual labor of creating metadata and free data generators and curators to focus on describing the potential uses of the data and data's particular caveats. Our project will experiment with large language models (LLMs) trained and fine-tuned on different texts, such as SciBERT, Llama 2, and GPT-4, to characterize their abilities to draft metadata for different datasets. We will create human-AI workflows for metadata creation and evaluation.

Data archives face challenges balancing privacy for individuals represented in data and analytic utility for data users. Recognizing and managing potentially disclosive and sensitive identifiers and responses in datasets is key to protecting privacy. We propose to investigate how computational tools, including GAI, can assist data producers and curators with disclosure risk review (DRR), the process of reviewing datasets for these data and planning approaches to mitigate their associated risks.

Project Work Plan

Our project requires two parallel tracks of research. First, in metadata drafting and evaluation, we examine how GAI can decrease the time it takes to draft metadata and increase data discoverability. We characterized curator workflows and data searchers' behaviors in our previous work and used those workflows to identify tasks that GAI 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. 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 GAI experiments to characterize GAI's abilities to augment these tasks:

Experiment 1: Drafting metadata descriptions. In this experiment, we will compare different LLM models' performance on a description 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 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.

Experiment 2: Generating summary statistics. In a second set of experiments, we will ask LLMs 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, 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.

The second track of research focuses on DRR and privacy protection. In this track, we will first define a taxonomy of GAI safety that indicates which GAI can be used to process the data. For instance, some sensitive data should not be analyzed on machines connected to the public internet, and they would be designated as “GAI safe” only when the model and data can be accessed on air-gapped computers. Using this taxonomy, we will experiment how GAI and other computational tools can enhance DRR and protect data privacy while balancing data utility needs.

Experiment 3: Enhancing DRR and suggesting privacy-friendly data use plan. We will test various GAI prompts and fine-tuning approaches to identify potentially sensitive datasets 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. Our recent work in dataset recommendation enables us to identify alternate datasets that may be more suitable for given research when appropriate risk remedies would be unworkable.

Personnel and Resources

Principle Investigator Libby Hemphill will commit one summer month each year to the project and will be responsible for managing project staff, securing appropriate computing resources, and setting the research agenda. We are requesting support for 2 UMSI PhD student research assistants for the first two years of the grant, 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. Our team also includes curators to be named from Deep Blue Data and ICPSR.

Diversity Plan

In aligning with the principles of our research plan, we are committed to not just dismantling barriers to data access and analysis, but also 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. 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.

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.

Budget Summary

The estimated budget is $749,460. IMLS Direct Costs include $96,780 Salaries/Benefits, $5500 for dissemination travel, $58,888 for Research Costs (including publications, cloud computing and storage, and curation services) and $347,813 Student Support Costs. (IMLS Direct Costs $508,981 + $240,479 IDC @ 56% = $749,460).