Viewfinder

Case Study

Newspaper collection metadata

Erin Wolfe, Digital Initiatives Librarian, University of Kansas

September 2026

DOI: 10.15788/1788283312

Project page for Viewfinder:
www.lib.montana.edu/responsible-ai/

Abstract

The University of Kansas Student Newspaper Collection is a 189k-page digitized microfilm collection that, until this project, lacked any meaningful metadata beyond broad, reel-level date ranges, making the collection difficult to search or discover. To address this, I used open-weight generative AI models to create detailed descriptive metadata and searchable Optical Character Recognition (OCR) text for the entire collection, producing nearly 1.9 million indexed content objects and close to 300 million words, with substantially improved reading order over the library's previous OCR tool. I used Viewfinder to reflect on this project from my own perspective as the project implementer, as well as from the perspectives of the researcher, library administrator, and library IT stakeholder roles. Working through multiple stakeholder perspectives helped me see values, tensions, and assumptions that were outside of my normal considerations and helped me gain a more complete perspective on this project.

Project details and background

The KU Student Newspaper Collection (SNC) at the University of Kansas (KU) consists of nearly 189k pages of images digitized from microfilm and converted into a digital collection at the KU Libraries. The SNC is the most heavily used microfilm collection in the KU Libraries Special Collections. However, it lacked any metadata or index beyond broad date ranges available at the reel level, making discoverability challenging and searchability impossible. The goal of this project was to increase access to the collection, facilitating findability and research opportunities. The digitization phase was itself a large project, with many KU Libraries faculty, staff, and students involved, including conservation review, communication with vendors, quality assurance of returned images, subject expertise, and more.

The AI-powered portion of this project—including image processing, script development, AI implementation, and data post-processing—was conducted by KU's Digital Initiatives Librarian. This phase used open-weight generative AI models to create a detailed index of the contents of each digitized page, along with descriptive metadata for each identified item. This was approached in a multi-step process, the first of which involved developing a custom Python script to query a hosted instance of GLM-4.5V, a vision-language thinking and reasoning model (106B parameters, with 12B active parameters per token). The submitted

query included the image itself, along with one or more steering prompts depending on page contents. To facilitate the large number of requests, this script was run in a distributed computing environment, with up to 50 instances running simultaneously.

Following the metadata generation phase, a similar approach was used to generate AI-enhanced OCR text via the GLM-OCR model, a lightweight vision-language document-parsing model (0.9B parameters, combining a CogViT vision encoder with a GLM language decoder). The output was saved as both plain text OCR and hOCR (hypertext markup language-based OCR) with bounding box coordinates at the content-block level.

The benefits of this approach at a large scale are significant, providing greatly increased access to the collection. For the 189k-page collection, an itemized page-level index containing nearly 1.9 million content objects was created, with descriptive metadata for each (including title, summary, keywords, category, named entities, and more), as well as nearly 300 million words (in 15 million content blocks). Testing the generated OCR against Tesseract output (the OCR generation tool that the library has used in the past) shows a slight improvement in word and character error rates (particularly for small, dense text), and significant improvement in reading order on the page, a particularly tricky issue for newspaper text (Kendall's tau: GLM-OCR 0.997 vs. Tesseract 0.734).

Data analysis and quality assurance are still in process, as is the final phase of this project, which involves creating an online portal to facilitate access to the data in a variety of ways. Although there is not really an AI component for the final phase (beyond some assistance with JavaScript and SQL coding), this piece connects the user with the data in a meaningful and granular way.

Viewfinder exercise

The Viewfinder exercise was conducted after the bulk of the AI implementation and processing was finished but midway through both the data analysis and the web access development. Following the first step in the exercise to identify relevant values from my perspective, the values that I identified as driving my approach were trust, service quality, and human-centeredness — all of which seem interrelated in this context.

The level of detailed metadata generated by this process is resource-intensive and would not be possible without the use of these AI tools. However, the potential for errors by metadata generated through these means is well-documented. In addition to normal human error, AI-generated output may include data that is misleading (e.g., an inaccurate article summary), incomplete (e.g., articles missed during the generated inventory), or inaccurate (e.g., hallucinations, incorrect names identified).

Given the complexity of the project and the volume of metadata generated, I used a human-in-the-loop approach to validate and appraise the various outputs throughout the process. As the person ultimately responsible for implementing and reviewing this AI-generated metadata, it is important to me that the data be as accurate as possible, while acknowledging the use and limitations of AI and drawing the user's attention to where errors may arise. Providing direct access to the original page image along with the metadata allows the user to verify their findings. This intentional transparency does not eliminate data uncertainty, especially when dealing with large amounts of results, but it does allow a ground truth source that can be referenced at any point.

The biggest downstream impact that I anticipate from this process is when the generated metadata becomes disconnected from its source and our own public-facing warnings about AI usage. The longer this metadata is public, the more it is likely to be scraped and reused in contexts where the existing built-in disclaimers won't travel with it, which may be more likely to be treated as ground truth by a user with no way to know its provenance or limitations.

In the second step of the exercise, which asks the participant to consider relevant values from a different stakeholder perspective, I first considered the role of researcher. This was really the easiest role to assume, because, although I had not undertaken a conscious process as with this exercise, I had been conducting this entire project with the end user in mind, with the goal of creating accurate, usable metadata. For the researcher's values, I highlighted trust and social responsibility, and added the value of usability.

Trust and usability were obvious choices. The researcher wants to be able to trust that what is presented to them is reliable, accurate, precise, and thorough. In addition, they want to be able to find the results that they are looking for in a user-friendly and expected way, and to be able to access the results in a format

that suits their needs. The results should be replicable and exportable to help support trust in their own output.

Regarding social responsibility, the researcher may want assurance that the metadata they're searching was generated with minimal environmental impact and is as free from bias as possible. Aspects of the model itself, such as inherent bias stemming from the model's training data, are outside of my control. However, I did track the model and release version used for all metadata generated, which provides documentation and a path for future review if concerns arise. I do recognize that provenance tracking is not the same as addressing environmental cost or bias directly, and this is something that should be given more attention for future projects.

Viewfinder challenged me to consider further perspectives, and I repeated the exercise with a different stakeholder role of library administrator. Taking on this role, I tried to stay focused on the AI-generated portion specifically, rather than the digitization project as a whole. Ultimately, I based these values on conversations and communications in my own institution, attempting to reflect what appears to be the emphasis of the current library administration, and settled on the values service quality, fiscal responsibility, and alignment.

Service quality here comes down to trust in the metadata itself and how that reflects on the library as a whole. The metadata was generated automatically, at a scale that is impractical to check in its entirety, so an administrator would reasonably ask whether it's safe to present the metadata as authoritative under the library's name. Misleading and inaccurate metadata is almost a certainty at this scale, and the library's role as a provider of trustworthy information needs to be considered. Previous library discussions at this level have concluded that archival metadata is often imperfect (i.e., errors in human-created metadata also exist), and the benefits in this case of such a highly increased level of access outweigh the existing lack of metadata at all, especially when combined with prominent and prevalent disclaimers about the source of the metadata.

Fiscal responsibility requires a large-scale project to consider both the benefits (e.g., usefulness of the output to patrons, potential for future application of skills and lessons learned) and the costs (e.g., financial costs of tools and services, human-hours, etc.). This type of project should consider if the library's effort is justified in relation to the final results. Library administration would be concerned with how this overall project, approach, and output align with departmental,

library, and university priorities. Is this helping the library meet its goals? In this case, the tools used were open-weight models that required no licensing or model-access costs. Although there were significant person-hours involved, KU librarians are actively encouraged to explore and use AI in a variety of ways, and this project helped to build skills and experience that can be applied to other collections, shared with colleagues, or used in assistance with researchers in the future.

After working through the previous stakeholder roles, I finally considered values that would be important from a Library IT perspective. I chose privacy and consent, alignment, and added the value technical sustainability. Of high importance to our IT is the security of KU's data, and specifically what happens to data once it's sent to an AI model (i.e., data handling) and potential downstream impacts on data subjects (e.g., named individuals found in the publication). On the surface, this isn't really an issue for this project, since the materials are all public, already-published works produced at KU and by KU students. All AI processing was done using models hosted on servers not connected to an internet-facing service, so no data was sent anywhere that could be leaked into a model's context or integrated into a training set (although the generated data is subsequently public itself, and almost certainly already scraped and folded into training data by other means).

When thinking about alignment and technical sustainability from an IT perspective, I found several perspectives: considering this project against existing technical and security policy for software and AI tools, as well as how the eventual web portal can coexist with existing software and handle web traffic (including heavy traffic from data-scraping bots). The distributed computing system that was used had already been pre-vetted by IT for different contexts, so it was not a concern for this project. IT could conceivably be concerned with why an open-weight model was chosen over a proprietary one for this kind of processing. My preliminary quality testing showed the chosen models to produce better results for a lower cost than proprietary models, which provided justification. As AI trends continue, a future in which AI models would need to be vetted by IT is possible, but that is not currently the case.

Discussion and outcomes

I found Viewfinder to be quite useful for prompting consideration beyond my own normal perspective. As noted, I had informally considered the role of the

researcher, but the Viewfinder exercise helped push that further. As I worked through the exercise, I found myself naturally noticing other perspectives that were not included, or values that were valid, but not among the three selected.

I did find that the further I moved from my own perspective, the more I had to work to retain focus on AI-specific concerns. My tendency was to frame the project more broadly, and it took a deliberate second pass to keep each value tied to the AI processing itself. This was probably the main practical lesson that I took from the exercise: it's easy to consider the most familiar or convenient meaning in a given role, rather than keeping the focus on AI-related ethics.

I also noticed some overlap and differences across roles that I found valuable. For example, I selected trust for both myself and the researcher. During implementation, I was concerned with accuracy and completeness, whereas the researcher needs to be able to rely on the output without independent verification. I relied on my testing, the selected model, and my ability to create effective prompts and quality check for errors; the researcher is (consciously or not) trusting in my process, decisions, and verification steps. This presents two ways of looking at the same thing, but with quite different emphases. On the other hand, trust and service quality can really emphasize the same underlying question, depending on the stakeholder's perspective. Recognizing these overlaps helped clarify that a single value can carry real differences in meaning, while two different values can evaluate the same concept from different angles. This distinction helped me properly frame where the tangible differences between perspectives actually lie.

For others pursuing similar work, I'd recommend running Viewfinder more than once, at different stages, rather than treating it as a one-time exercise. Also, especially for larger projects with an AI component, I found it important to be deliberate about defining when a value's meaning quietly shifts from "about the AI" to "about the project." As an additional exercise, it could be informative to intentionally look for a stakeholder whose interests may conflict with the project's, not just ones who use or operate it.