Using Metadata as Data for Reparative Archival Description: A Case Study

Elizabeth James

Assistant Professor/Archivist and Digital Preservation Librarian

Libraries and Online Learning

jamese@marshall.edu

Dean: Monica Brooks

monica.brooks@marshall.edu

Abstract

The goal of this project was to conduct a case study in which one mid-sized archive, Marshall University Archives and Special Collections, could explore the issues inherent in approaching reparative archival description from a metadata as data standpoint. The project examined automated ways of evaluating existing metadata describing archival materials and proposed a model for addressing problematic metadata through actions that take place during archival work, including the creation of collecting policies, appraisal, processing, description (both free-text description and controlled vocabularies), and digitization. The project culminated in the creation of a website sharing results to encourage reproducibility and reuse by other institutions: https://elizabethjames.net/s/mad-rad.

Introduction

Archivists and librarians rely on classification systems, controlled vocabularies, and processing approaches that prioritize certain perspectives while erasing others. As a result, it is critical that the field examine metadata creation and descriptive practices frequently, critically, and iteratively using all techniques at our disposal. However, archives are chronically under resourced financially and in terms of adequate staffing levels. The drastic disparity between available resources and the scope of archival work results in several well documented issues: increasing backlogs of archival materials that are not described and thus inaccessible to researchers, inadequate or culturally insensitive metadata describing digital materials, and high levels of stress and burnout among archivists. The second issue is one that has increasingly come to the fore of the archival profession in light of the events of summer 2020, which included widespread protests in response to the murders of George Floyd and Breonna Taylor that further surfaced issues of structural systemic racism in the United States. In response, archivists began to interrogate structural and systemic white supremacy within the profession and the institutions in which archivists work. At that point in time, with the majority of Marshall staff and faculty working remotely in varying conditions that exposed additional issues of inequity such as lack of quality internet access, stable housing, childcare, and food instability that inordinately affected communities of color, archivists discussed some of the systemic issues that we were unthinkingly perpetuating by not addressing legacy issues affecting equitable and ethical access to archival materials. However, with limited staff and no funding or approval for additional personnel, a manual review of digitized materials was not feasible. Automated and open-source tools,

however, offered a path in which archivists could conduct this much needed work within existing institutional constraints.

Results

As part of this project, I conducted an analysis of metadata describing 8,202 items that are part of Marshall University's manuscript collections using AntConc and the appropriately named Topic Modeling Tool, based on MALLET, a topic modeling package. Before delving into the employee-created data that is more subject to bias, such as an item's title, free-text description, and relevant subject headings to examine the specific descriptive choices, I realized that any analysis would be remiss if it did not include a more global overview that examined what items were digitized, when the items were created, and from what collection the items originated. This is because every item and collection that was digitized and described was once chosen for digitization for by the same individuals who demonstrated bias in description. As such, I was unsurprised to find that certain collections and formats were overwhelmingly more represented than others. The items represented were more visual in nature with 88.5% of materials consisting of photographs, postcards, and drawings. Visual materials are often scantly labeled and require more subjective description by the person creating metadata for the materials. Three collections, out of ninety collections total, supplied 32.3% of the items digitized and available online. With my experience as an archivist working with these collections, I knew that they documented two distinct areas: 1.) original greeting card art that featured often generic imagery and 2.) local genealogical and historical subjects that almost exclusively documented white families, white owned businesses, and other topics that neglected to include a fuller scope of items related to the history of the area though the premise of the collection would have, in theory, included more diverse histories.

Once I had a sense of the limited scope of the collections I would be examining, I understood that I would be unlikely to identify some of the issues I sought to address in the project proposal. I was already well acquainted with using AntConc and Topic Modeling Tool, and usage and analysis of the results output by these tools was relatively straightforward, though the actual results output surprised me. Given the initial scope of the data used based on my early analysis using Excel, I did not expect to find many areas in which I would be able to address some of the problems I initially identified as being the focus of this project because the data I

used did not represent diverse topics. However, despite being unable to identify any issues regarding description of marginalized races and identities due to the lack of representation of these subjects and individuals in the collection, I was able to find significant equity issues regarding the representation of women that I would have been unable to do had I not used a data-centric approach to analyzing these materials. In AntConc, the issue I identified was that women were often only identified by their husband's name, even when only cursory research would have been able to easily find a name to identify these women. As a result, women were stripped of agency and identity and subsumed by their husband's name and identity instead. When using the Topic Modeling Tool, I had issues when working with what I termed as "microtexts" of metadata records. These records, even when examined at scale, did not make the same meaningful collections they did when working with longer-form texts. The topics discovered, in most cases, simply replicated the subject matter of collections with large amounts of digitized materials.

An additional issue that occurred was a significant increase in scholarship on this issue throughout the summer of 2021. The quantity and frequency of literature forthcoming made a meaningful and current literature review not only difficult, but impossible to accomplish in a meaningful way. As a result, the project had to pivot to emphasize an approach that focused on ways of determining possible interventions at the systemic and institutional level that seek to address what might be understood as archival "algorithms of oppression," a term coined by Safia Noble in her book of the same name, that seeks to provide a term for the systematic issues that perpetuate marginalization of minority populations. The approach that I took as part of this work, one that utilized metadata describing digitized collections as data, has proven to be a novel one not used elsewhere even with the surfeit of literature that was produced throughout spring and summer 2021. Using these tools with an emphasis on creating a replicable project that prioritized transparency required me to articulate principles and define approaches that make the work of identifying and updating problematic metadata less prone to bias by a particular archive or archivist in comparison to manual approaches that have been the focus of reparative metadata projects in the field.

However, such data-centric work had an unintended but unsurprising consequence: it put the lack of diversity of our collecting materials on display by surfacing what was most represented in our digital collections. As a result, I was able to broaden the potential impact of

the project to encompass the idea of reparative work as not only being a process of intervening and editing existing metadata in accordance with culturally competent description, but also on how archivists determine what to collect and how they prioritize what materials are digitized and thus made more widely available to a larger audience. Reparative work requires acknowledging and seeing the absences and gaps in archival collections as much as what is present. As a result, the website examined available tools and automated approaches of using reparative principles to evaluate existing metadata describing archival materials and propose a model for addressing problematic metadata through actions that take place during archival work, including the creation of collecting policies, appraisal, processing, description (both free-text description and controlled vocabularies), and actions taken as part of the digitization process.

Discussion

Working on this project made apparent that automated approaches must be complemented by thoughtful examination of the results output by the tools by an experienced and culturally competent archivist. Careful and considered archivist interpretation and interrogation of not only the results, but also how the data is used for analysis, can help curtail the perpetuation of systemic issues. Systemic issues can be perpetuated if the archivist fails to interpret the output in a way that acknowledges how the overwhelming frequency of the majority of information can eclipse the minority of records. Without a thoughtful approach, analysis with automated tools risks reinforcing the same oppressive generalities that it seeks to address. However, an awareness of these issues and the desire and ability to identify them using the automated means and tools identified as part of this project proved to only address part of the larger issue. Despite the focused nature of the project on identifying and remediating existing problematic metadata, the results of the project proved to have significantly wider ramifications for archival work. Ultimately, the final products created, consisting of a model of ways that bias-aware practices can be incorporated into archival work and a set of recommendations for specific areas of archival practice, demonstrate that a commitment to conducting reparative work that seeks to truly succeed requires universal rather than localized intervention. Because of this project, archives will have a framework for examining and addressing ways in which bias can creep into all archival functions and address systemic issues before they make it to the stage of public description. Through these changes in operations, archives will be able to create collections that document the breadth of historic and contemporary events, issues, people, and

organizations. As a result, the archival record will be one that is more complete and less exclusionary in comparison to archives in the past, which primarily sought to document the perspectives of white, higher class, men. As a result, archives will be able to mend relationships with communities it has traditionally ignored. These more complete archives will be better able to connect with the concerns and interests of contemporary researchers, meeting user needs and breaking cycles of erasure of marginalized voices.

Appendices

Since the project resulted in the creation of a public-facing website that emphasized transparency in every aspect, any materials that contributed meaningfully to the project may be found on the site. The URL for the site can be found here: https://elizabethjames.net/s/mad-rad.