

Overview

Working with the digitized cards from the David M. Rubenstein Rare Book and Manuscript Library's physical card catalogs, our team explored the **50,000+ cards** to further the library's initiative of finding and describing historically marginalized voices in their collections.



Fig 1. Former card catalog cabinet of drawers

Data Wrangling Methodology

Internet Archive Search Tool	This tool allows keyword search and drawer exploration of the cards. Links to individual cards are found in the dataset.
OCR	Using Google Tesseract, we reran optical character recognition on the jpegs to convert images to usable text.
Collecting Author Names	Using regex and NLP part of speech tagging, we pulled out some of the author names. Then, we manually corrected the errors in OpenRefine.
Sorting by Collection	Created a scoring algorithm that sorted cards as a collection header or narrative using a combination of recurring patterns including length, POS tags, and keywords.
Metadata Extraction	Using regex for year and SpaCy for entity recognition, we were able to collect many dates and locations, but were limited by OCR quality.
Data Analysis	Using Python with Pandas in Jupyter Notebook, we completed exploratory, geospatial, demographical, and historical analysis on the dataset.
Data Compilation	Using Streamlit, we compiled our data and analysis into an interactive web app for easy viewing of our results. All files will be uploaded to Duke RDR.

Card Catalog Demographics

Most of the collections are **from North Carolina, Virginia, and surrounding states**. The bulk of the collections are from the United States, but there are also many from other countries, Europe and Asia in particular.

NC County Card Catalog Frequency

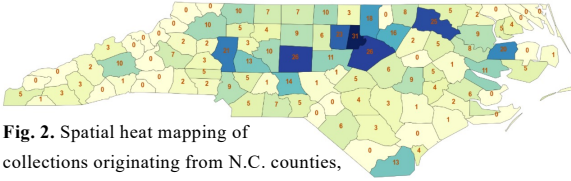


Fig. 2. Spatial heat mapping of collections originating from N.C. counties, many clustered around Durham County.

About 20% of collections were authored by organizations. Of those by people, **most were created by men**.

Gender Ratio of Collection Authors

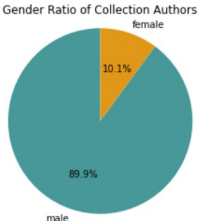


Fig. 3. Gender ratio of collection authors.

Limitations

- Even using a top OCR software, errors were plentiful, and compounded through subsequent steps of the project.
- The catalog is a static resource, and we can only use the information that was added to it when it was in use

Final Deliverables

- Internet Archive Search Tool¹
- Structured Dataset²
- Interactive Web App³

Common Topic Insights

1. Common topics found in the data include the **Civil War, family, politics, business, religion, and foreign affairs**
2. Duke's Early Names, Presidents, faculty, building names
3. History of Slavery, Charleston Earthquake, Wilmington Race riot, Labor movements

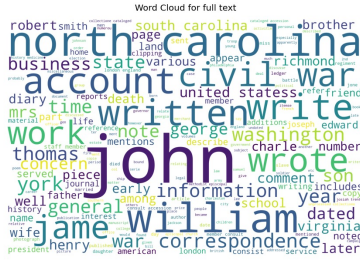


Fig. 4. Word cloud with common terms of full text of dataset, common categorical words removed for clarity.

Future Directions

- Additional manual data correction
- Identification of “outdated language”
- Further historical research: Methodism, Civil War, activism in NC
- Sentiment analysis of groups: slaves, southern gentlemen, southern belles

Acknowledgements and References

We would like to thank our project manager Anna Holleman, project lead Meghan Lyon, Paul Bendich and Gregory Herschlag at Data+, and Eric Monson from the Duke Center for Data and Visualization Sciences for their support.

¹See the Internet Archive: <https://archive.org/details/rubensteinmanuscriptcatalog>

²See the data in the Duke RDR: <https://doi.org/10.7924/r4br8v905>

³See the web application: <https://share.streamlit.io/bini-a/rlapp/main/app.py>