

USCT Pension Files Project

Executive Summary

Client

Dr. Edda Fields-Black

Student Consulting Team

Amiya Chokhawala (achokhaw)

Angela Kim (eunseokk)

Cathy Wang (cathyw)

Felicia Zhang (feliciaz)

Background

The USCT Pension Files Project, led by Dr. Edda L. Fields-Black at Carnegie Mellon University, is a research initiative focused on recovering and making accessible the family histories of United States Colored Troops (USCT) veterans and their communities. The project supports historians, genealogists, descendant communities, and public history institutions by transforming handwritten Civil War pension files into searchable historical resources. A central goal of the project is to overcome the “1870 Brick Wall,” which has historically limited genealogical research for African American families prior to the 1870 Census.

Project Description

Project Opportunity

The USCT Pension Files archive contains approximately 95,000 handwritten pension files with valuable genealogical information such as names, family relationships, locations, and personal testimony. Prior to this project, extracting information from these records relied primarily on manual transcription, which was accurate but not scalable. At the same time, high transcription accuracy is essential because even small errors can significantly impact genealogical interpretation. The project opportunity was to design a structured workflow that improves scalability while preserving the level of accuracy required for historical research and public access.

Project Vision

The project vision was to support a hybrid workflow that combines AI-assisted transcription with human validation and structured data organization. Instead of relying entirely on manual transcription or fully automated extraction, the team developed a modular pipeline that integrates automated grouping, community verification, and public-facing exploration tools. This approach creates a foundation for transforming handwritten pension files into reliable and searchable historical data that can support long-term research and descendant access.

Project Outcomes

The project produced three primary outcomes across technology, process, and people dimensions.

Technologically, the team developed a prototype pipeline consisting of an automated entity grouping system, a Zooniverse-based community verification workflow, and a public search

interface prototype. Together, these components demonstrate how transcription data can be structured, reviewed, and presented through an integrated workflow.

From a process perspective, the project introduced a scalable verification workflow that replaces informal transcription coordination with a structured crowdsourcing model. Documentation of pipeline scripts, validation workflows, and grouping logic supports continued development by Dietrich Computing and future student teams.

From a people perspective, the system expands participation in transcription validation by enabling historians, genealogists, and descendant communities to contribute corrections through an accessible interface. Early user testing confirmed that volunteers can successfully review transcription content when provided with structured context and highlighting support.

Project Deliverables

The project deliverables include a working prototype entity grouping pipeline, a Zooniverse-based transcription validation workflow, and a public search interface demonstration. Additional deliverables include documentation of the verification process, user testing materials, pipeline scripts, and repository resources designed to support future scaling and integration with the broader USCT Pension Files infrastructure.

Recommendations

To support long-term sustainability, the report recommends expanding Zooniverse testing to larger volunteer groups, implementing reviewer consensus scoring methods, developing automated postprocessing for volunteer corrections, and integrating validation outputs into the entity grouping pipeline. Additional recommendations include strengthening partnerships with genealogical organizations and collaborating with Zooniverse developers to improve transcription interface features such as formatting support and uncertainty tagging. Together, these steps provide a pathway for scaling the prototype into a sustainable workflow for community-supported historical transcription validation.

Student Consulting Team

Angela Kim worked on the Zooniverse community verification workflow and coordinated stakeholder communication and user testing. She is a senior majoring in Information Systems and is interested in roles that connect technical systems with user needs.

Amiya Chokhawala worked on designing and testing the pension file exploratory interface and coordinated client communication. She is a junior in Information Systems with minors in AI and Business. She will be working as a product management intern at Visa this summer.

Felicia Zhang designed the core entity grouping scripts while managing technical alignment with Dietrich Computing. A junior in Information Systems and Statistics/Data Science, Felicia is an incoming Software Engineering intern at Google with a strong interest in data science.

Cathy Wang worked on the Zooniverse setup pipeline and postprocessing support. She is a junior majoring in Information Systems, with an additional major in AI. She is interested in roles related to software, AI engineering and machine learning research.