# Digitizing a Lifetime of Genealogy: Creating a Searchable NLM Dataset

CS4991 Capstone Report, 2025

Riley Immel
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia USA
immel@virginia.edu

**ABSTRACT**
Far away in Wyoming is a tiny attic closet full of papers constituting all that is left of Charles Luxmoore's life-long work tracing his family history back to the ninth century. As his grandchild, I want to not only preserve his work but to expand upon it by creating an app that enables members of our family to search the extensive family tree using natural language. I propose utilizing software to convert the scans of physical documents into digitized versions from which a database containing all the people and family ties can be built. Further, I propose utilizing a natural language model and artificial intelligence to build an interface enabling people to ask questions about the data, instead of writing formal structured searches. Once this has been accomplished, I intend to expand the project by finding ways to integrate it with other genealogical works and moving it beyond my grandfather's extensive work.

## 1. INTRODUCTION

The preservation of historical records is an essential task for maintaining a connection to our past. In my family's case, a lifetime of genealogical research—carefully documented by my grandfather, Charles Luxmoore—has been confined to a collection of physical papers stored in a dusty attic closet. Converting these documents into a digital format not only safeguards the information against physical deterioration but also opens new possibilities for accessing and exploring our family history.

To address the challenges involved in preservation, my project proposes a twofold solution. First, optical character recognition (OCR) software will be employed to convert scanned images of handwritten and printed documents into text. This text will then serve as the basis for constructing a structured, searchable genealogical database. Second, rather than relying on traditional, structured queries, the project will integrate a natural language model to allow family members to interact with the database conversationally. In the following sections, I detail the technical approach and review prior work in the fields of document digitization and natural language query interfaces that have informed this project.

## 2. RELATED WORKS

A key component of this project is the extraction of text from historical documents, for OCR is essential. The Tesseract OCR engine has been widely adopted in digitization efforts. Smith (2007) provides an in-depth overview of Tesseract's architecture and performance in converting printed text into machine-readable form. This work has proven influential in numerous digitization projects and offers a solid technical foundation for processing genealogical records.

In tandem with OCR, enabling intuitive, natural language interaction with complex datasets is crucial. Jurafsky and Martin (2009) offer a comprehensive introduction to NLP, detailing methods for parsing, semantic analysis, and query interpretation. Their text lays the theoretical groundwork for systems that translate everyday language into structured database queries. Additionally, the practical capabilities of NLP have been further advanced by toolkits such as the Stanford CoreNLP (Manning, et al., 2014), which demonstrates how robust language processing can facilitate user-friendly interfaces in a variety of applications.

Large-scale digitization initiatives also offer valuable insights. For example, the National Library of Medicine has made significant strides in preserving and providing access to historical documents through its digital collections. Similarly, FamilySearch has transformed the accessibility of genealogical records by integrating digital archival practices with searchable databases. These projects illustrate the broader context and benefits of digitization, reinforcing the approach taken in this project.

## 3. PROCESS DESIGN

To explain the design of the project, I have broken it into its linear steps, each explained below, starting with a brief whole project overview.

### 3.1 System Architecture Overview

The proposed system is designed as a modular pipeline that integrates three major components: Document Digitization, Data Parsing & Database Construction, and a Natural Language Query Interface. Each module is developed to function both independently and as part of an integrated workflow, ensuring that improvements or changes in one component do not disrupt the overall system. The modules can be thought of as building blocks that can be chained together to form an overall structure or be used independently for their singular purpose.

### 3.2 Document Digitization and OCR Conversion

The first stage involves scanning the physical documents containing historical genealogical data. These high-resolution scans will then be processed using the Tesseract OCR engine (Smith, 2007) to convert the images into machine-readable text. Prior to OCR processing, image enhancement techniques such as noise reduction, contrast adjustment, cropping, and aligning are to be applied to improve text recognition accuracy. The output will be segmented into discrete records that represent individual entries within the family tree.

### 3.3 Data Parsing and Database Construction

Once the raw text is obtained via OCR, custom parsing routines will extract key data elements such as full names, birth and death dates, and familial relationships such as spouse and children. This information is then normalized and mapped onto a structured relational database schema tailored to represent the hierarchical nature of genealogical information. Data cleaning procedures address inconsistencies—such as variable date formats, varied name spellings, or incomplete records—to ensure a high-quality, searchable dataset. Once the database is filled with the baseline information, construction of the family tree(s) can begin and the vast array of connections can be built out into the database.

### 3.4 Natural Language Query Interface

To make the vast repository of genealogical data accessible to users without technical expertise, the system integrates a natural language processing (NLP) interface. Leveraging principles outlined by Jurafsky and Martin (2009) and implemented with

toolkits such as Stanford CoreNLP (Manning et al., 2014), this component interprets everyday language queries. User input is parsed and translated into structured database queries, allowing users to retrieve information intuitively, whether they are asking for direct relationships or more complex historical patterns. The overall goal is to allow queries that not only are very akin to traditional database queries such as "Who is X's father/third son?" but also less formal queries such as "Who in X's family immigrated to America in the 1800s?"

### 3.5 Integration and Workflow

The complete workflow is designed to ensure a seamless transition from physical documents to a digital, searchable database. The pipeline starts with document scanning and OCR conversion, followed by data extraction and normalization, and finally culminates in an interactive query interface. This modular architecture not only facilitates iterative development and testing but also allows for future scalability—such as integrating additional data sources or enhancing query capabilities with more advanced NLP models.

## 4. ANTICIPATED OUTCOMES

Given the early stage of the project, the following anticipated results have been identified based on preliminary experiments and design simulations:

Preliminary tests on a small subset of documents have shown that applying image enhancement techniques before OCR processing significantly improves text recognition accuracy. Initial experiments indicate that, after appropriate preprocessing, the Tesseract OCR engine can reliably extract key textual data from the scanned historical documents. Although exact accuracy metrics will be refined as the system scales, early indications suggest that the integration of these methods will substantially reduce transcription errors compared to manual data entry as well as compared to conversion with no preprocessing.

In addition, the natural language query interface is expected to greatly enhance accessibility. By translating user queries into structured database requests, the system aims to reduce the learning curve associated with traditional search methods. Anticipated outcomes include faster retrieval times, improved user satisfaction, and the ability to handle more complex and less formalized queries, as family members will be able to navigate complex genealogical relationships using familiar language. Ultimately, these improvements are projected to lower both the time and effort required to access detailed historical records, paving the way for broader use and potential integration with other genealogical datasets.

Another anticipated outcome is to create a product that allows for conversational manner, almost as if the questions were being asked of my late grandfather himself. With generative AI and NLP in their current state of development and the progress they've made paired with personal experience using ChatGPT to ask questions conversationally, it is expected that the desired outcome of having a low-barrier-to-entry product will be achievable in the case of this project.

## 5. CONCLUSION

My proposed solution aims to lower the barrier for users who are not familiar with formal query languages, thereby enhancing accessibility. The integration of OCR and natural language processing (NLP) is significant for several reasons. Digitization using robust OCR techniques preserves the integrity of historical documents while simultaneously creating opportunities for innovative data interaction. By allowing users to ask questions in everyday language, the

project democratizes access to complex genealogical information and encourages deeper exploration of family relationships and history. Ultimately, this work aspires not only to preserve an invaluable personal archive but also to serve as a model for future projects involving the digital transformation of historical records.

## 6. FUTURE WORK

Although preliminary experiments have demonstrated promising results, several enhancements are planned for the next phases of the project. Future work will focus on further refining the OCR accuracy by integrating advanced machine learning techniques for image preprocessing and character recognition. Additionally, further work needs to be done in terms of cropping, editing, and categorizing the scans to help improve the OCR performance.

Once results from the OCR stage have begun to consistently meet the desired quality standards, the project can begin to move on to the next phase which is data parsing and building the database. Once the overall project has fully gotten underway, I plan to spend considerable time on choosing how to implement the NLP module to ensure the best quality and remove the need to backtrack later. Finally, a long-term goal is to create a mobile application version of the project to expand the portability and accessibility of the project.

## 7. ACKNOWLEDGMENTS

I want to first express my thanks and gratitude to my late grandfather, Charles E. Luxemoore for the literal decades of his life he spent collecting, compiling, researching, and building the family trees and books of genealogical data upon which this project is built. Without his work, I would know far less about my family history and this project would not exist. I want to also thank my mother and brother, Carolyn and Kyle, for the time they spent with me scanning the physical copies of my grandfather's work. Last, I want to thank my professors here at the University of Virginia for giving me the knowledge and skills to be able to take on a project like this.

## REFERENCES

Smith, R. (2007). An overview of the Tesseract OCR engine. In Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) (Vol. 2, pp. 629–633). IEEE.

Jurafsky, D., & Martin, J. (2009). Speech and language processing: An introduction to natural language processing, Computational Linguistics, and Speech Recognition. Prentice-Hall.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations (pp. 55–60). ACL.

National Library of Medicine (NLM). (n.d.). Digital Collections. Retrieved from https://collections.nlm.nih.gov/

FamilySearch. (n.d.). Family History Library. Retrieved from https://www.familysearch.org/en/