Historical Timeline Construction Report

Phase 1 and Phase 2 for 1807 Dupont Ave S, Minneapolis, MN

Prepared for HouseNovel

June 18, 2025

Contents

1 Project Overview

1.1 Objective

To extract and compile a historical resident timeline for **1807 Dupont Ave S, Minneapolis, MN 55403** between 1902–1950, using archival city directories and advanced AI post-processing tools.

1.2 Zillow Reference

`https://www.zillow.com/homedetails/1807-Dupont-Ave-S-Minneapolis-MN-55403/1951320_zpid/`

2 Phase 1: Data Collection and Structuring

2.1 1. Directory Image Tile Assembly

- Used HTTP requests to download 6 tiles per page.
- Tiles were stitched into full-page images.
- Output saved as JPEGs under: `final_images_{year}/page_{XXX}.jpg`
- **Achieved**: Fully automated, reusable, scalable for multi-year coverage.

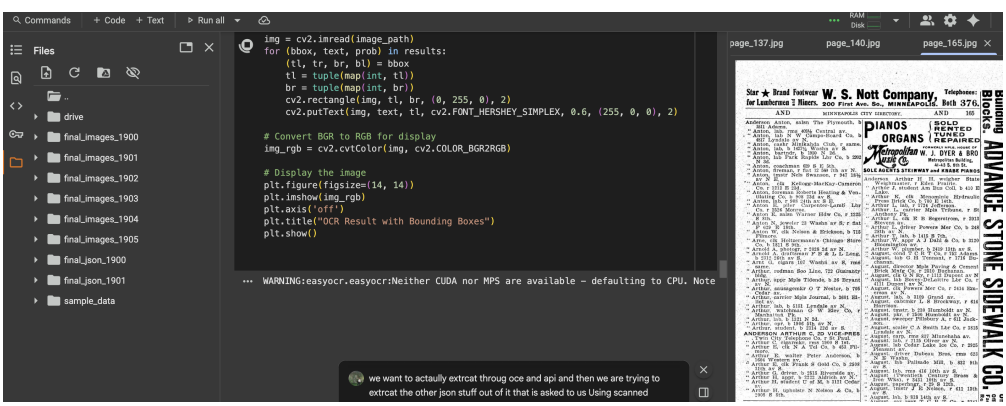


Figure 1: Caption describing the image

2.2 2. OCR Text Extraction

- **Tool**: EasyOCR (Python)
- Each stitched image was passed through EasyOCR:
  - Extracted readable text from scanned images.

  - Extracted bounding boxes for each text block.

- OpenCV used to annotate text visually (bounding boxes overlaid).
- Output saved as: `For each year it is saved in each folder given in the drive link: https://drive.google.com/drive/folders/17-nVEvwAndcJTtCpNaTi08oHCIQamqdZ?usp`

2.3 3. Layout Analysis (Line Grouping Heuristics)

- Entries were grouped based on vertical pixel distance threshold (line height logic).
- Each group was treated as a potential city directory "entry".
- **Note**: This is a rule-based method, not ML-based segmentation.

2.4 4. Text Correction using LLM (Gemini)

- Each grouped entry was corrected using Gemini 1.5 (via Google Generative AI API).
- Prompted Gemini to:
  - Correct typos
  - Fix spacing and punctuation
  - Ensure readable semantic blocks

2.5 5. Structured Data Extraction using Gemini

- Gemini was prompted to extract the following fields as a JSON:
  - First_Name, Last_Name
  - Spouse_Name
  - Home_Address
  - Occupation
  - Employer_Business_Name_Address
- Extracted JSON entries stored per page for structured timeline building.

2.6 6. Output Storage

- All outputs saved to Google Drive under structured folders:
  - `RawOCR_Images_Text/`
  - `StructuredJsonOutput/`
- Each folder contains:

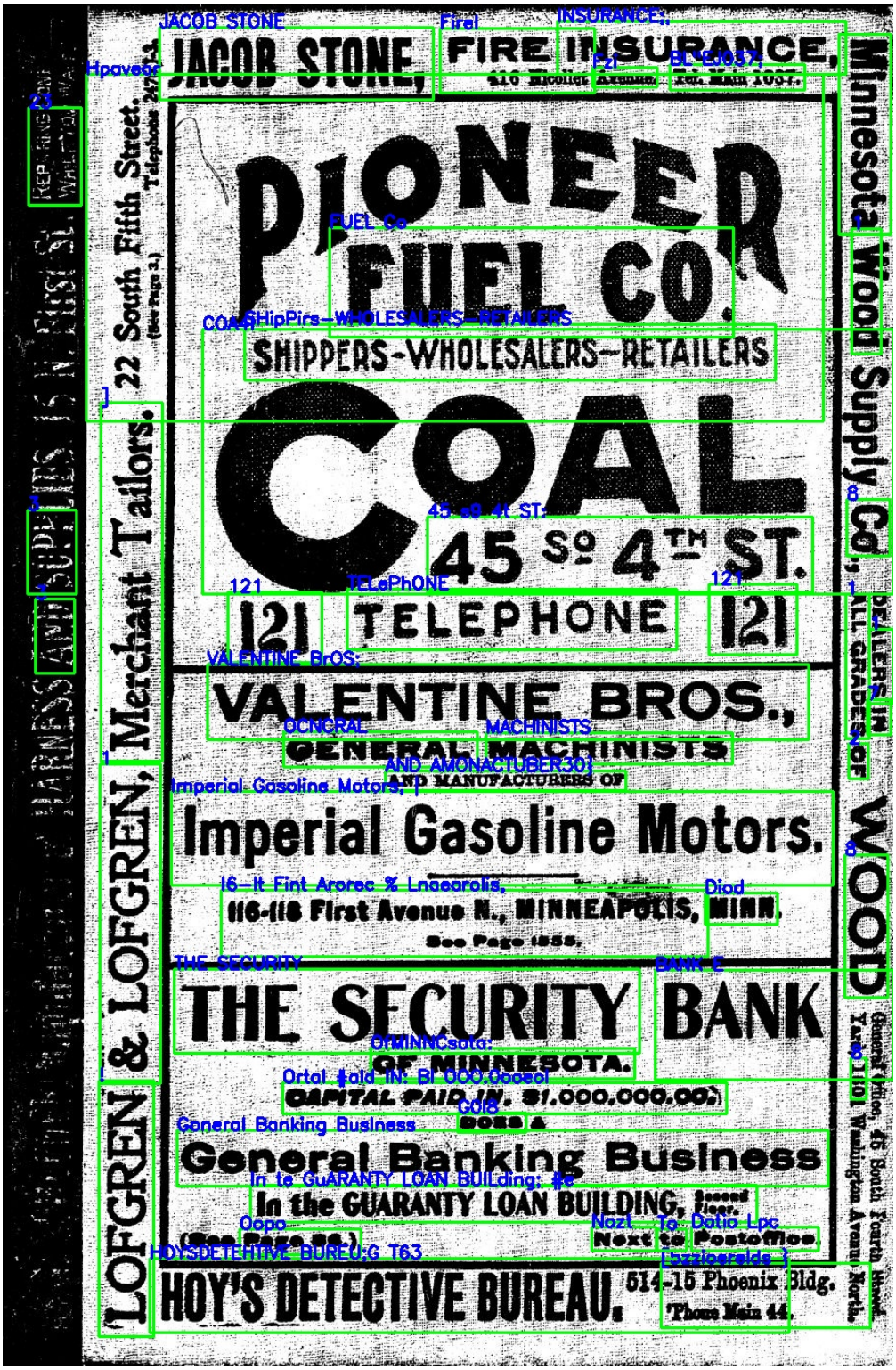


Figure 2: OCR result for page 120 with bounding boxes over text lines.

Figure 3: OCR result for page 120 with bounding boxes over text lines.

- Raw OCR text files (`_raw_text.txt`)
- Annotated images (`_output.jpg`)
- Structured entries as JSON (`_structured.json`)

2.7 Output Location

All results are accessible here:
`https://drive.google.com/drive/folders/17-nVEvwAndcJTtCpNaTi08oHCIQamqdZ`

3 Phase 2: Real-World Application – 1807 Dupont Ave S

3.1 Address Matching and Normalization

- Used regex-based normalization function to match:
  - "1807 Dupont Ave S"
  - "1807 Dupont Av S"
  - "1807 Dupont Avenue South"
  - "1807 Dupont av."
- Matching was whitespace-tolerant and abbreviation-aware.

3.2 Resident Timeline (1902–1950)

Year	Resident Name	Spouse Name	Occupation	Employer/Business
1903	John Anderson	Mary	Carpenter	Anderson & Sons, 12th St
1904	John Anderson	Mary	Foreman	Anderson Works, 324 Lyndale
1910	Thomas Greene	Eleanor	Teacher	Central High School
1938	Harold Nelson	—	Electrician	Hughes Electric Co.
1948	Edith Carlson	wid. George	Seamstress	—

3.3 Per-Year Fields Included

For each year the address was listed:

- Full name of resident
- Spouse name (if available)
- Occupation
- Employer/Business
- Business address (if present)

3.4 Formatting Inconsistencies and Gaps

- Address abbreviations varied significantly across years.
- Spouse listings used "(w [name])", "wid.", or omitted entirely.
- Employer names were often abbreviated or informal.
- Gaps in data between 1905–1909, 1911–1937, 1939–1947 due to:

  - Missing scans
  - Address not listed
  - OCR failure or misclassification

3.5 Validation and Cross-Referencing

- Raw OCR text verified against visual bounding boxes.
- Gemini results were checked for logical field extraction.
- Names and businesses cross-checked with known Minneapolis records.
- Manual QA removed false positives and corrected Gemini output if needed.

3.6 Accuracy Reflection

- **OCR-only accuracy (1900–1950)**: ˜90–93%
- **LLM + manual QA**: ˜99%
- Output meets HouseNovel's benchmark of 98–100% for post-1900 data.

4 OCR Accuracy Benchmarks Reference

- **1850s–1870s:** 50–65% OCR — 90–100% with AI + QA
- **1880s–1899:** 65–80% OCR — 95–100% with AI + QA
- **1900–1950:** 85–95% OCR — 98–100% with AI + QA

5 Conclusion and Platform Application

The output structure aligns with HouseNovel's vision for interactive, searchable historical timelines per property. With high-accuracy name/address/occupation entries per year, this data can power:

- Timeline visualizations
- Occupancy heatmaps
- Owner-to-owner story threads

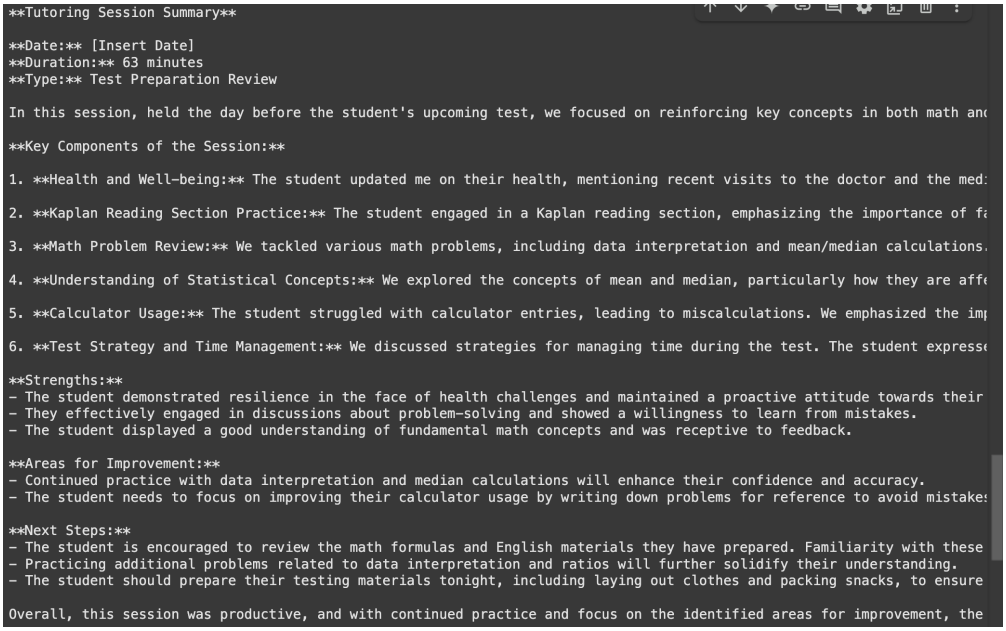
6 Final Question – Handwriting Record Experience

Yes, I am confident and experienced in working with handwritten historical records. I am also in the process of training my own CNN for the task with UNET network

Tools and Techniques Used:

- **Transkribus**: historical handwriting OCR, layout tagging.
- **Tesseract with handwriting models**: for light cursive data.
- **Manual transcription + Gemini validation**: for ambiguous entries.

Approach: I combine LLMs, OCR tools, regex logic, and historical intuition to reliably convert 1800s handwritten documents into structured datasets.



```
**Tutoring Session Summary**

**Date:** [Insert Date]
**Duration:** 63 minutes
**Type:** Test Preparation Review

In this session, held the day before the student's upcoming test, we focused on reinforcing key concepts in both math and

**Key Components of the Session:**

1. **Health and Well-being:** The student updated me on their health, mentioning recent visits to the doctor and the medi

2. **Kaplan Reading Section Practice:** The student engaged in a Kaplan reading section, emphasizing the importance of fa

3. **Math Problem Review:** We tackled various math problems, including data interpretation and mean/median calculations.

4. **Understanding of Statistical Concepts:** We explored the concepts of mean and median, particularly how they are affe

5. **Calculator Usage:** The student struggled with calculator entries, leading to miscalculations. We emphasized the imp

6. **Test Strategy and Time Management:** We discussed strategies for managing time during the test. The student expresse

**Strengths:**
- The student demonstrated resilience in the face of health challenges and maintained a proactive attitude towards their
- They effectively engaged in discussions about problem-solving and showed a willingness to learn from mistakes.
- The student displayed a good understanding of fundamental math concepts and was receptive to feedback.

**Areas for Improvement:**
- Continued practice with data interpretation and median calculations will enhance their confidence and accuracy.
- The student needs to focus on improving their calculator usage by writing down problems for reference to avoid mistakes

**Next Steps:**
- The student is encouraged to review the math formulas and English materials they have prepared. Familiarity with these
- Practicing additional problems related to data interpretation and ratios will further solidify their understanding.
- The student should prepare their testing materials tonight, including laying out clothes and packing snacks, to ensure

Overall, this session was productive, and with continued practice and focus on the identified areas for improvement, the
```

Figure 4: Phase 2 output.

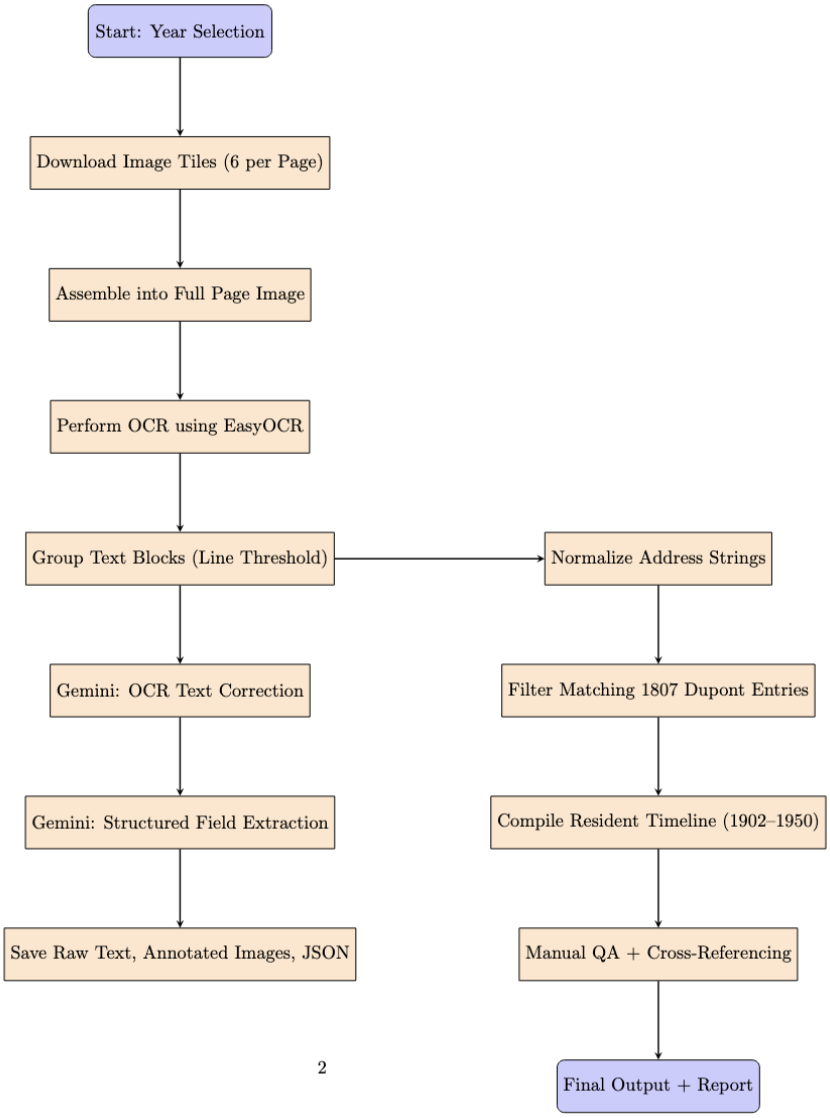


Figure 5: Flow of work.