Enriching Historical Records: An OCR and AI-Driven Approach for Database Integration

Zahra Abedi

Supervisors: Richard van Dijk & Gijs Wijnholds

Universiteit Leiden
The Netherlands

Index

Background

Linking University, City and Diversity (LUCD)

- Visualize interactions between Leiden University and city of Leiden since 1575.
- Capture the impact of students and professors from outside the Netherlands on Leiden.
- Collaborative work between researchers and students from LIACS and humanities faculty.
- A software system is designed which contains:
  - Database
  - Adapters for data extraction, transformation, loading and linking
  - Website for visualizing the results

Research Questions

Focus: Enriching the centralized database with “Leidse hoogleraren en lectoren 1575-1815” dataset.

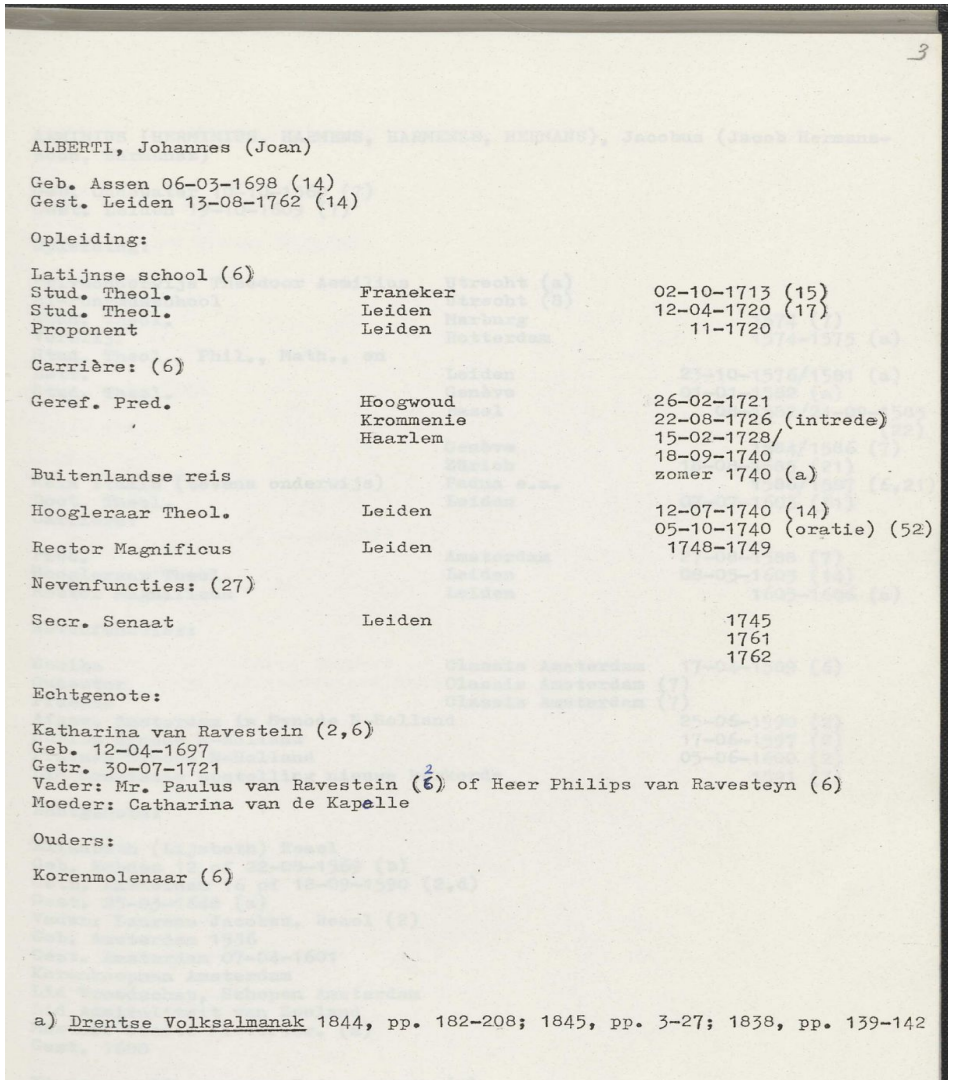
Research Questions: How can we accurately extract and transform historical records data from scanned historical documents and map it into a centralized database?

- SQ1: How can we extract high-accuracy text from scanned historical documents using OCR techniques?
- SQ2: How can AI play a role to analyze the OCR generated text and obtain a structured format?
- SQ3: How can we map the structured data into a centralized database?

Data

Leidse hoogleraren en lectoren 1575-1815

- Compiled by A.A. Bantjes and L. van Poelgeest from 1983 to 1985
- Seven volumes
- Contains the following information about professors:
  - Date and place of birth and death
  - Education
  - Career history
  - Additional positions
  - Genealogical details regarding: Spouse(s), Children, Parents, Grandparents, etc.
  - Special details (salary, memberships, etc.)
- Printing quality varies among different volumes



3

ALBERTI, Johannes (Joan)

Geb. Assen 06-03-1698 (14)
Gest. Leiden 13-08-1762 (14)

Opleiding:

Latijnse school (6)		
Stud. Theol.	Franeker	02-10-1713 (15)
Stud. Theol.	Leiden	12-04-1720 (17)
Proponent	Leiden	11-1720
Carrière: (6)		
Geref. Pred.	Hoogwoud	26-02-1721
	Krommenie	22-08-1726 (intrede)
	Haarlem	15-02-1728/ 18-09-1740
Buitenlandse reis		zomer 1740 (a)
Hoogleraar Theol.	Leiden	12-07-1740 (14) 05-10-1740 (oratie) (52)
Rector Magnificus	Leiden	1748-1749
Nevenfuncties: (27)		
Secr. Senaat	Leiden	1745 1761 1762

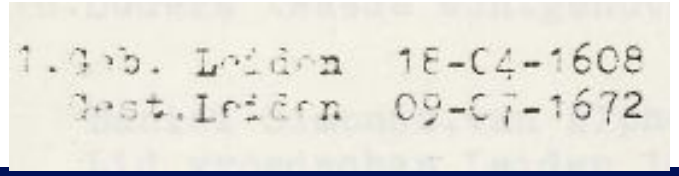
Echtgenote:

Katharina van Ravestein (2,6)
Geb. 12-04-1697
Getr. 30-07-1721
Vader: Mr. Paulus van Ravestein (6) of Heer Philips van Ravesteyn (6)
Moeder: Catharina van de Kapelle

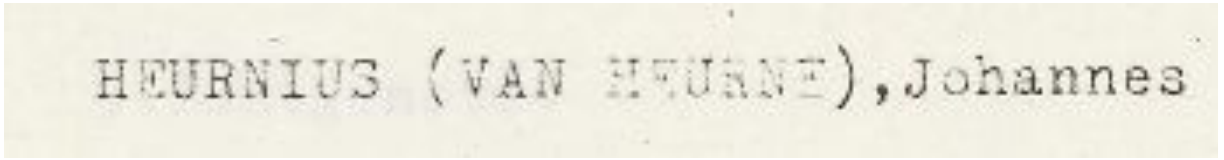
Ouders:

Korenmolenaar (6)

a) Drentse Volksalmanak 1844, pp. 182-208; 1845, pp. 3-27; 1838, pp. 139-142



1. Geb. Leiden 18-04-1608
Gest. Leiden 09-07-1672



HEURNIUS (VAN HEURNE), Johannes

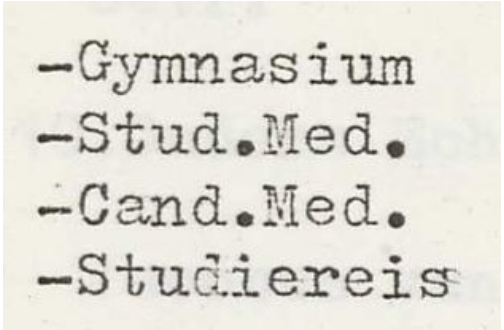
Methods

1. PDF to PNG Conversion

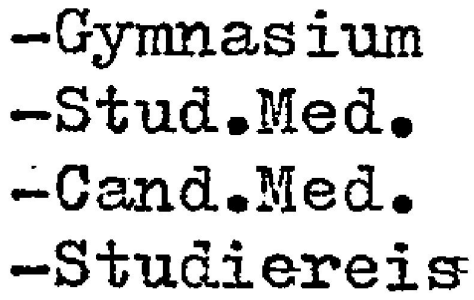
- Choice for PNG due to it's lossless image quality.

2. Image Preprocessing

- Using OpenCV (cv2):
  - OpenCV is a popular computer vision library for image preprocessing tasks.
- Image Denoising:
  - Reduces noise and artifacts present in the PNGs.
- Conversion to Grayscale:
  - Removes color information, leaving only shades of gray.
- Grayscale to Binary Conversion:
  - Converts grayscale images to binary format, where each pixel is either black or white.



Before Processing

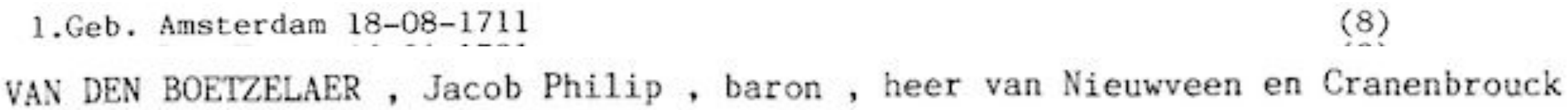


After Processing

Methods

3. OCR

- Extract the text from the images for further processing
- Tool Selection: Tesseract
  - Choice for Tesseract OCR engine due to its robustness and versatility in handling various font styles and languages.
- Training Tesseract:
  - Trained Tesseract using a a training set of image-text pairs from all seven volumes of the dataset.
  - Improve recognition accuracy and adapt to specific font styles and historical language nuances.
- Word List Integration:
  - Introduced a word list containing the most common words found in the dataset.
  - Serves as a reference for Tesseract during the OCR process.



Methods

4. Split text

- Split the text per person using Regular Expressions (regex) in Python.
  - Last names are typically written in all capital letters. Example: ALBERTI, Johannes (Joan)
  - Looking for a string with more than 3 subsequent capital letters.
- Challenges:
  - Handling multi-page information.
  - Regex performance may vary due to OCR errors. Example: “FAS” vs. “F45”
- Results:
  - Generated individual text files for each professor.
  - Some instances of two professors’ data combined due to OCR inaccuracies.

```
JUNIUS (DE JONGHE), Hadrianus (Adriaen)

Geb. Hoorn 01-07-1511 (14)
Gest. Arnemuiden of Middelburg 16-06-1575 (2,14)

Opleiding: (2)

Stud. Med. Bologna 1540
Doct. Med.

Carrière: (2)

Verblijf Parijs

Verblijf Engeland

Gouverneur Kroonprins

van Denemarken 1562-1563
Stadsgeneesheer Haarlem 1563

Rector Lat. school Haarlem 1563
Stadsgeneesheer Middelburg 1574
Hoogleraar Theol. Leiden 1575 (beroepen)

N.B. NOOIT AANGENOMEN
```

Methods

5. Large Language Models

- Function Calling for Converting OCR-generated text into a structured, valid JSON format for database integration.
- Model Selection: GPT 3.5
  - Function Calling: Ensures the AI consistently generates valid JSON outputs according to predefined schema.
- Challenges:
  - AI limitations in consistently extracting all relevant information from complex historical texts.
- Results:
  - Successfully generated structured JSON data, however refinement is ongoing to improve consistency.

Methods

6. JSON to Database Mapping:

- Field Mapping: Map the fields from the JSON schema to the corresponding fields in the database.

7. Database Modifications:

- Adapt the database to encapsulate all relevant data fields extracted from the JSON format by adding new tables and columns.

8. Data Linking:

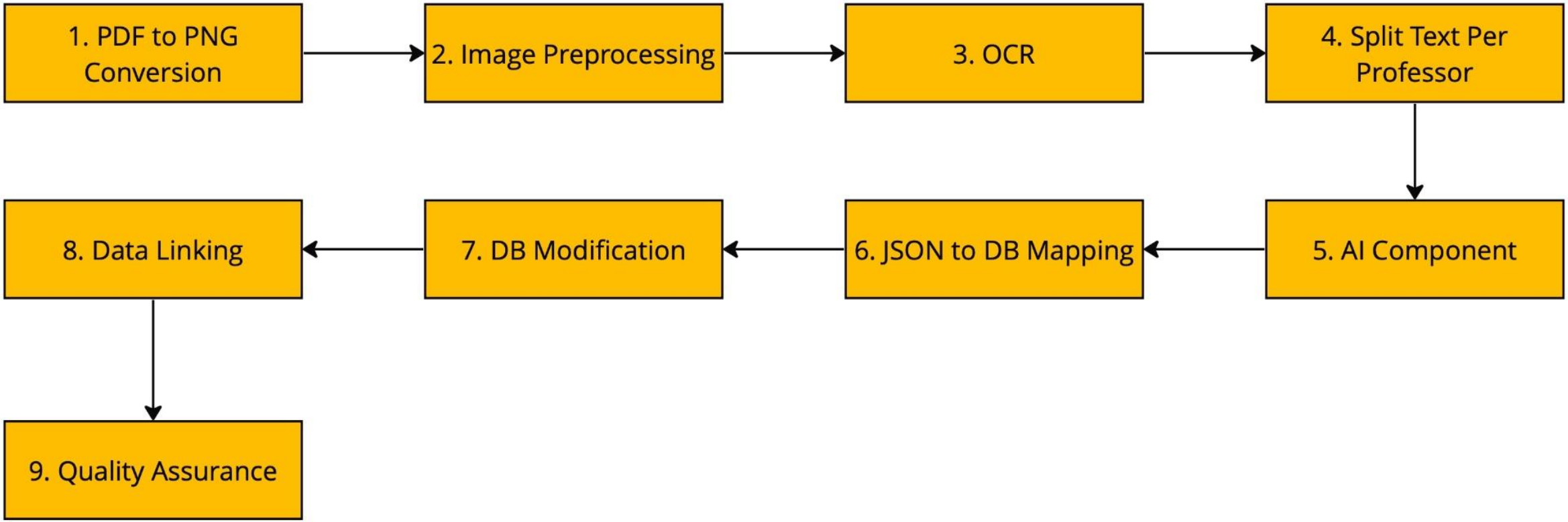
- Matching Criteria: Link instances based on first name, last name, birth place and birth year.

Methods

9. Quality Control

- Historian Verification:
  - All current database entries are checked by a historian to ensure accuracy.
- Data Rating System:
  - Existing Data: Higher rating assigned to pre-existing data.
  - New Data: Lower rating for newly added data until verified.
  - Post-Verification: Ratings can be adjusted to a higher value after historian verification.

Process Workflow



Thank You!

Universiteit
Leiden
The Netherlands

Discover the world at Leiden University