Zeng Family Tree

May 23, 2025

Note: Don't tell mom. She doesn't like these books. Maybe don't tell dad either until we got something cool to show

TODO: Create GitHub repo

1) Scan book

Status: book 1 scanned: book1

Scan all pages of all 4 books into PDF. Ideally use a nice photocopier instead of iPhone scans for high quality

2) Digitize book

Given the scans, we want all the information in the book to be digitized. For the trees, we want a tree structure with the Unicode characters for the Chinese characters as the nodes. For the biography entries, we want a string.

Approach 1: Hardcoded parsing

Step 1: Page scans -> Tree of Chinese character images

Type 1: Lineage tree

Status: script

We want to create a tree data structure from the page scans. The nodes (chinese characters) will still remain images at this stage.

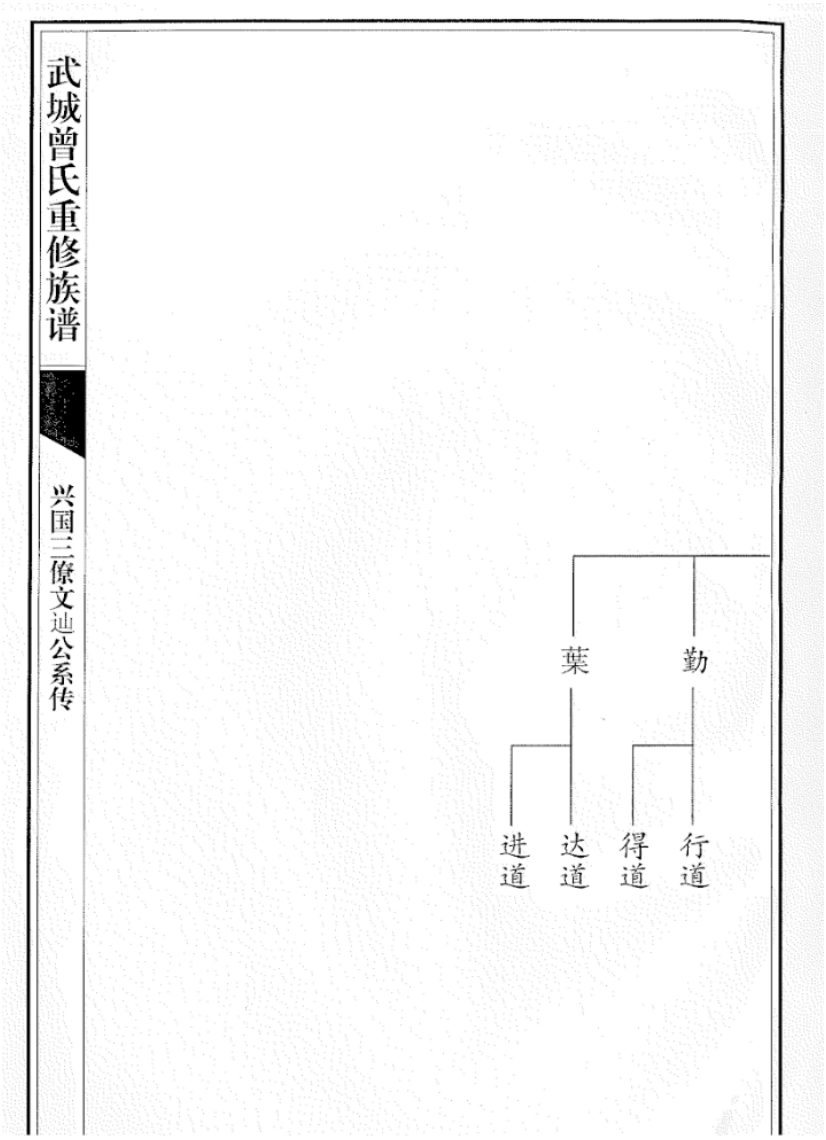
Challenge1: Branches sometimes travel between pages.

- One way to handle this is to have the ends of these branches be marked as UNK1/2/3, and join them between pages; if they don't connect with subsequent pages, throw an error.

Challenge 2: How to join the tree between different pages. One tree may start with 子, so we have to find the previous 子 node to connect it with.

- May be easier to connect once we OCR and know with more confidence which are identical characters. We could also just do an image similarity comparison
- Usually the previous 子 node should be a page or two ago, but idk for sure

For example, this page:

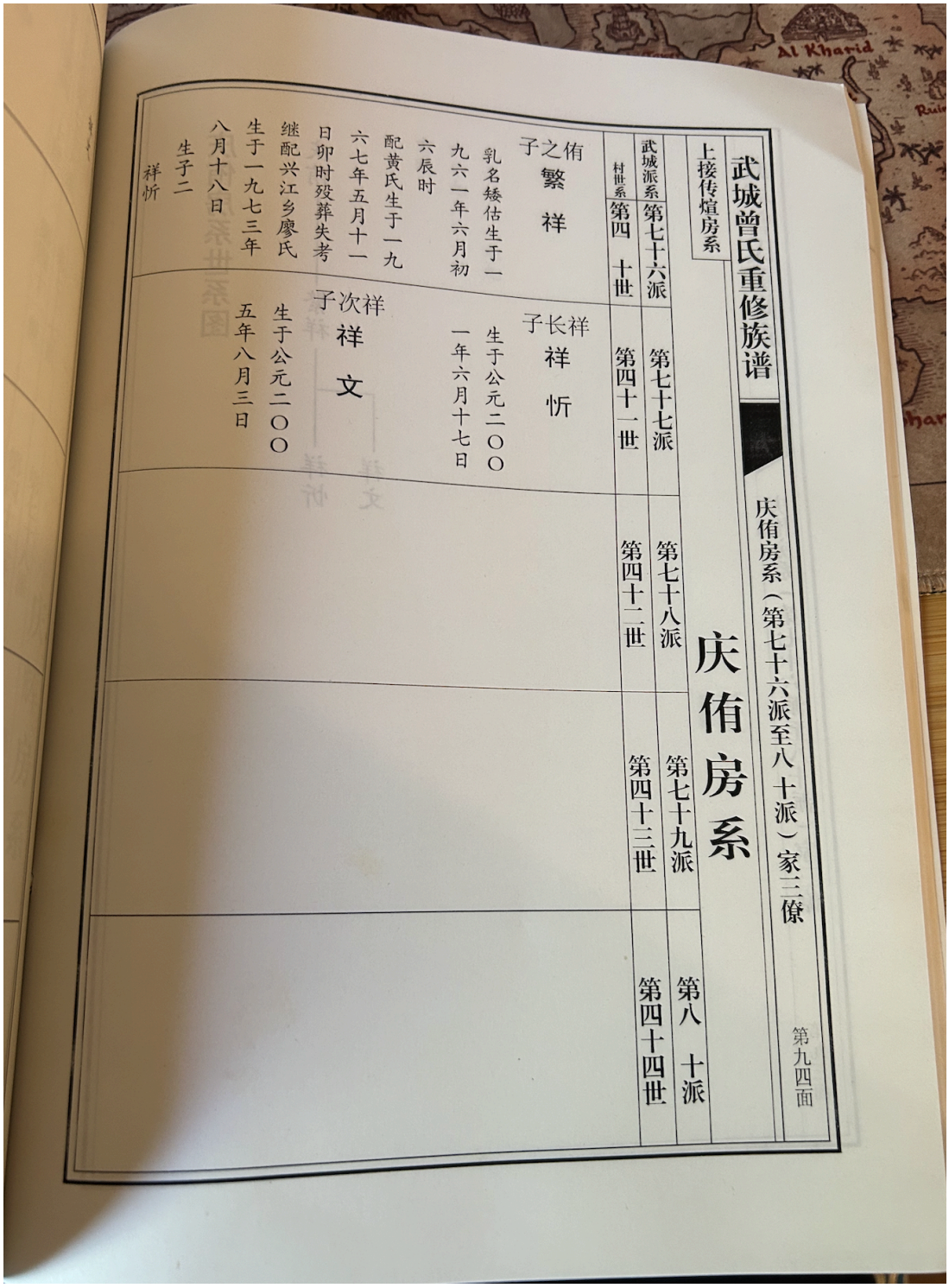


Should become:

(For two-character names, each character should ideally be separated into its own image, to make later parsing easier. Since all text in the books are the same font, the size of each image (i.e. its bounding box) should be identical, which we can hardcode.

Type 2: Biography

Looks like this



In books 3/4, these are interspersed with tree pages. We need separate parsing for these pages. Note the text is right to left, top to bottom. We also need to properly associate the biography with the nodes in the tree.

Step 2: OCR

Here, we want to take the images of all the Chinese characters from step 1 and convert them to Unicode characters.

Challenge: OCR sometimes doesn't detect some characters as they're ancient and not used in modern Chinese

Could try Google Document AI OCR, Mistral OCR, or something well suited for Chinese specifically. I think Mistral may be better than Google since it's more recent?

Since we have the images of all Chinese characters from Step 1, we can pack them densely into a page if OCR charges by page. We could use grid lines to separate the characters to make identification easier. This way, we also know exactly where all the characters are, and thus what their bounding boxes are. This way, we know if a character isn't detected.

For characters that aren't detected, we can manually convert them. Maybe Dad's down to help

Approach 2: Use AI

We could feed the pages into an LLM like Claude, and ask it to parse the image for us. May be expensive and potentially unreliable if it fails. Could give it a try though

3) Create static website

- Could start even after approach 1 step 1, and just use the images of the characters instead of their unicode text
- Could build something local first, then deploy live
- TODO: What domain?
- Could add search functionality once we OCR the text
- Could have an about page for Zeng generational names
- Language support for both Chinese + English?
- For famous zengs in our tree, could link to their wikipedia page or other articles about them
- Could have an about page with other famous zengs too outside our tree?

4) Create dynamic website

- Could have users for each of the Zengs on the site
- Could update biography entries, and keep version history tracking
- Online db?
- A user is an admin for all changes to their subtree (Ex. only us/dad/grandpa could update our node entries)

- Give the Zeng women biography entries too?