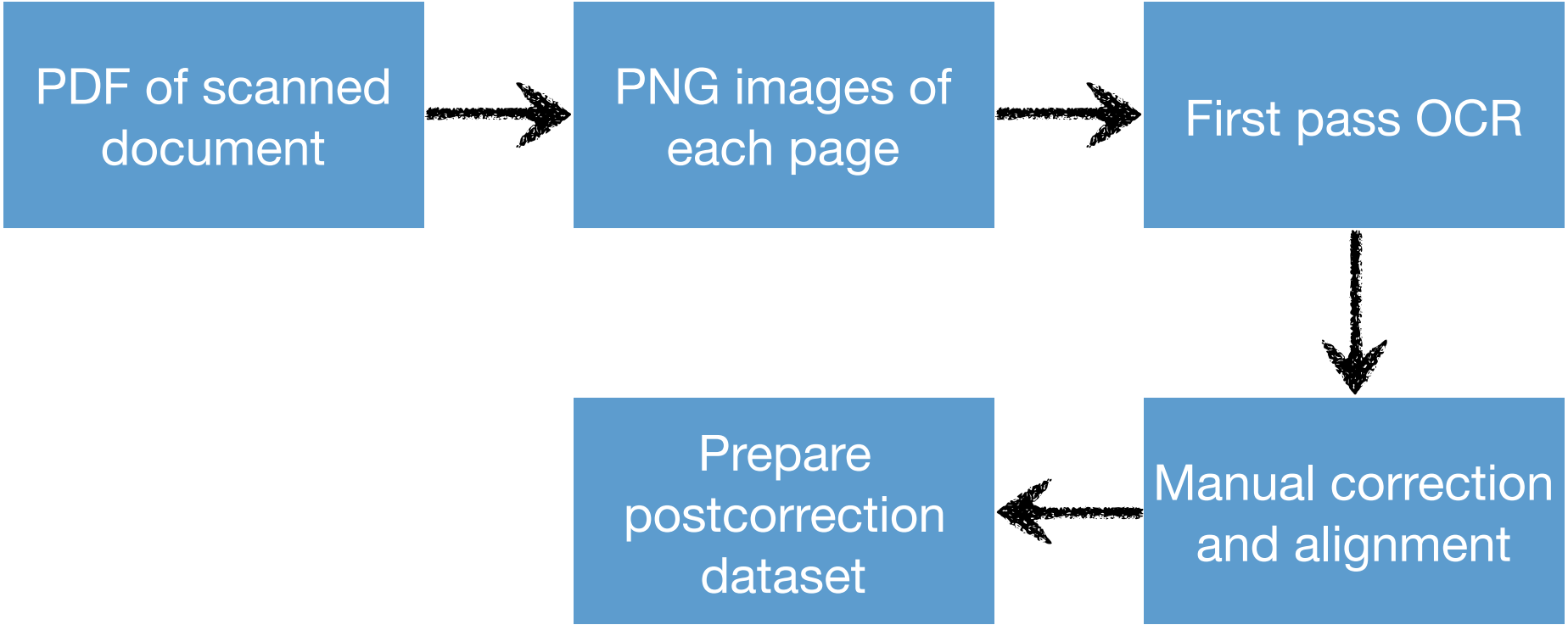
PDF of scanned document
PNG images of each page
First pass OCR
Manual correction and alignment
Prepare postcorrection dataset