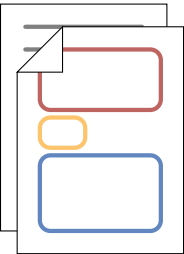
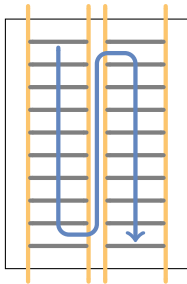


Document Sources



Step 1: determine which elements to extract.

Pre-processing



Step 2: Prepare xml documents by determining margins and adjusting text flow.

Element Classification



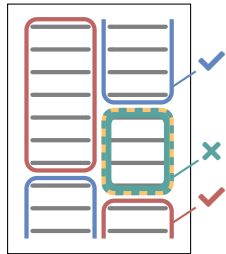
Step 3: Classification of text elements with the help of an active learning model.

Post-processing



Step 4: Improve element classification by separating documents and merging page breaks.

Validation



Step 5: Validate all aspects of the extraction process through manual assessment.