HJDataset

A Large Dataset of **H**istorical **J**apanese **D**ocuments with Complex Layouts

Zejiang Shen, Kaixuan Zhang, Melissa Dell

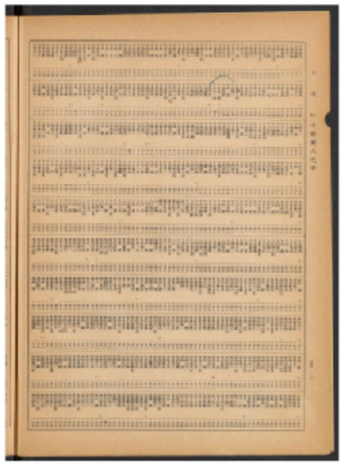
{zejiang shen, kaixuanzhang, melissadell}@fas.harvard.edu

HARVARD
Faculty of Arts and Sciences

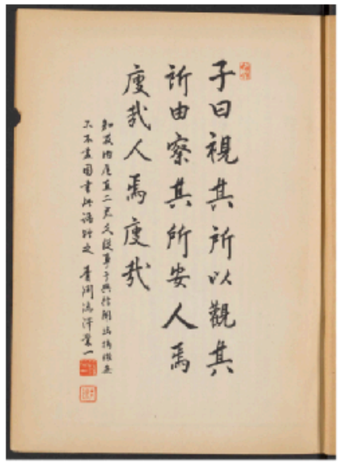
The Institute for
Quantitative Social Science

INHOMOGENEOUS PAGE SCANS

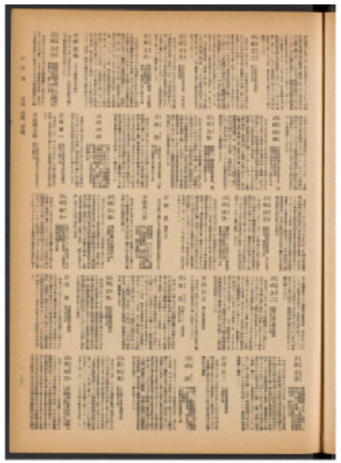
(a) Advertisement Page



(b) Index Page



(c) Other Page



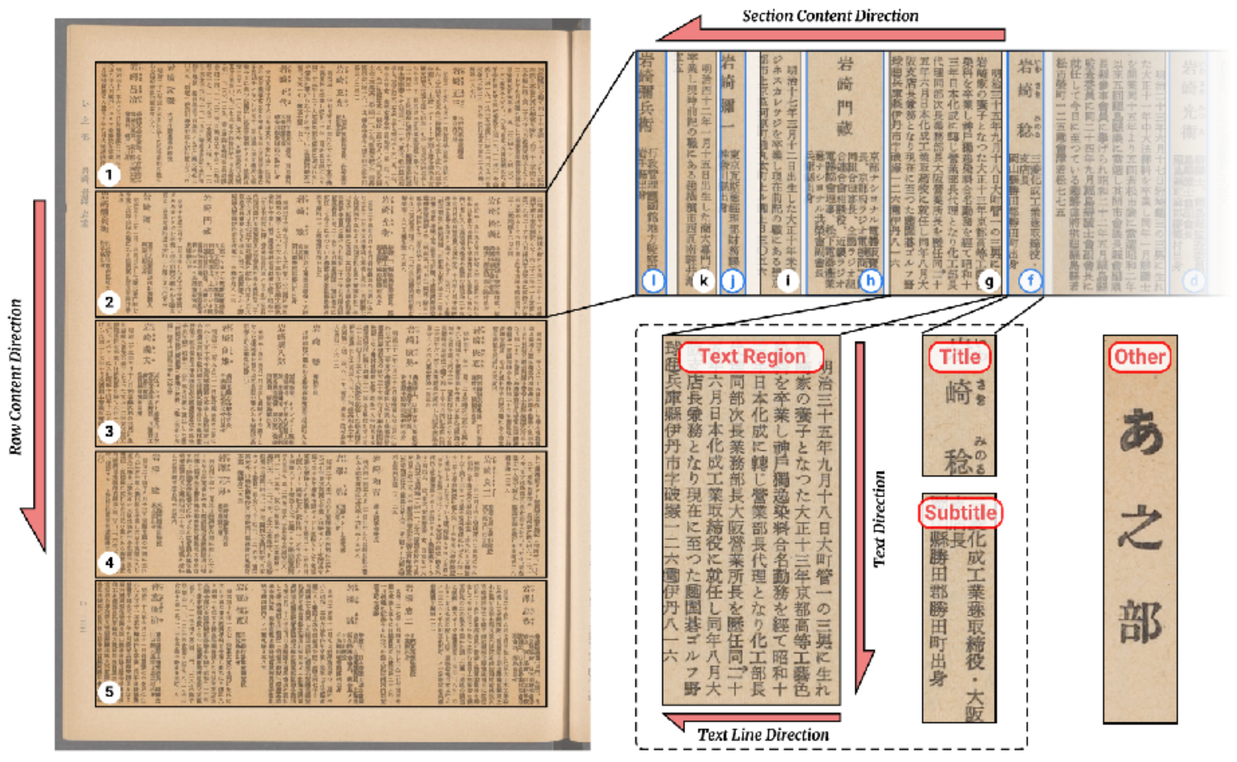
(d) Main Page

(e) Simplified Annotation Illustration

Whole Page | *Page Frame* *Row* *Text Region* *Title Region* *Title* *Subtitle* *Other*

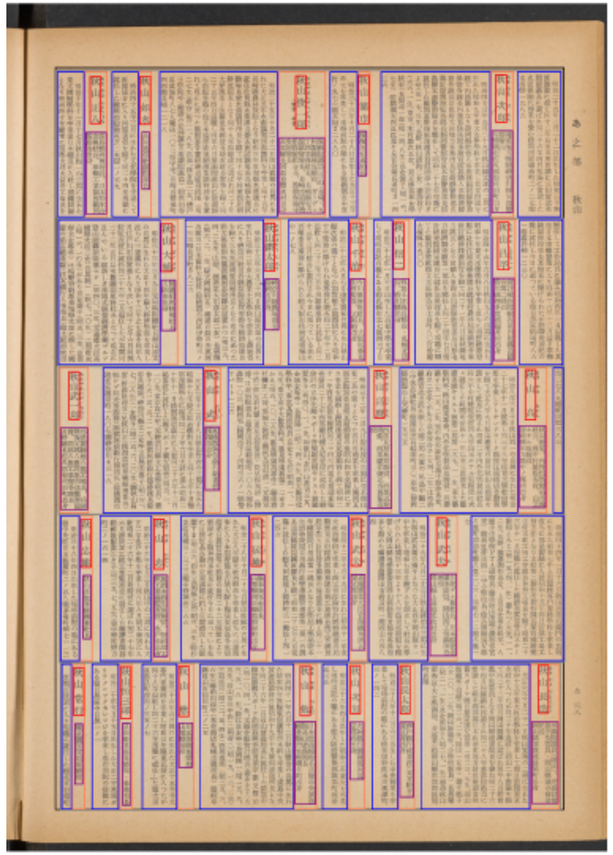
- HJDataset contains 2k page scans of four categories, as illustrated from (a) to (d).
- We provide 250k layout annotations for the index page and main page, and (e) is a simplified example for the layout structures.
- There are seven types of layout elements, and reading order and hierarchical relationships are included in the annotations.

Complex Page Layout Structure

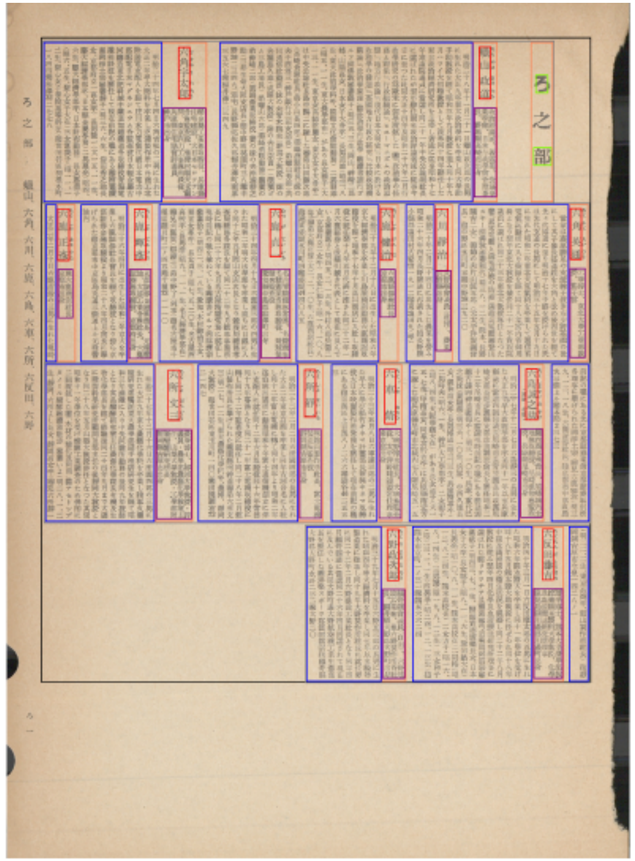


- For pages in the main book, they each contain five rows that are vertically stacked, and the text regions are horizontally arranged within each row.
- Texts are vertically written inside *text regions*, e.g. (g) in the figure.
- The *title region*, e.g. (f), can be further split into *title blocks* and *subtitle blocks*.
- An *other* category is reserved for chapter headers and other irrelevant text regions.

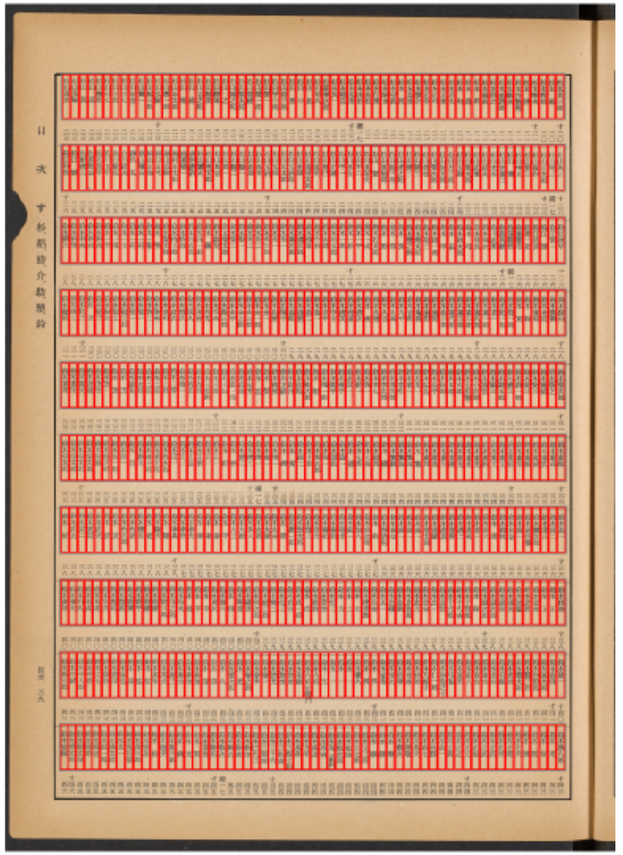
Page Layout Examples



(a) Main Page Annotation Example 1

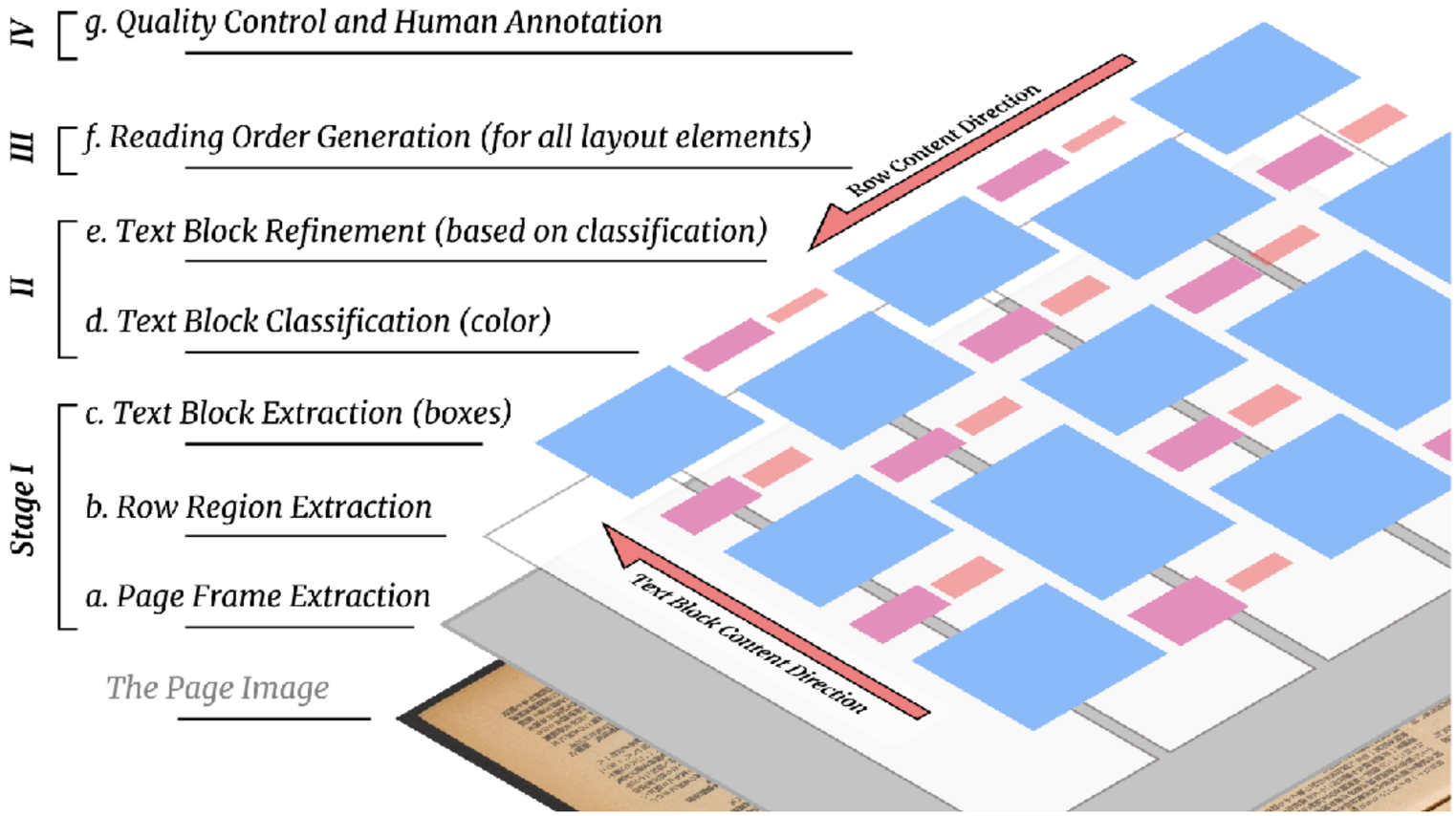


(b) Main Page Annotation Example 2



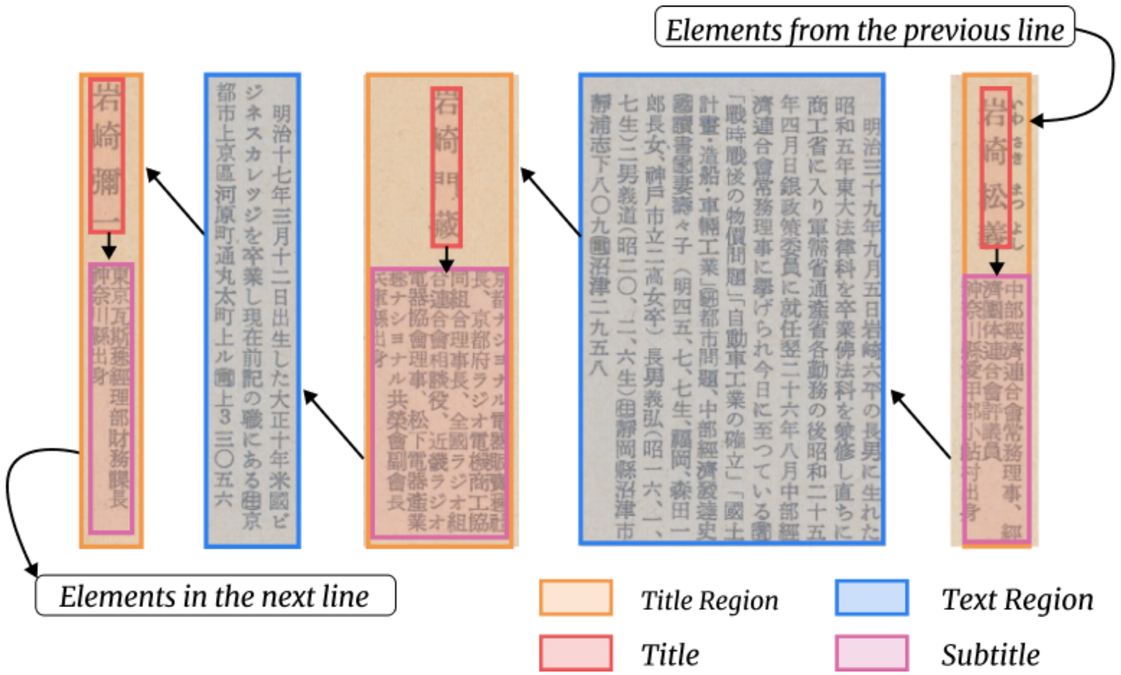
(c) Index Page Annotation Example

Semi-Rule Based Layout Generation

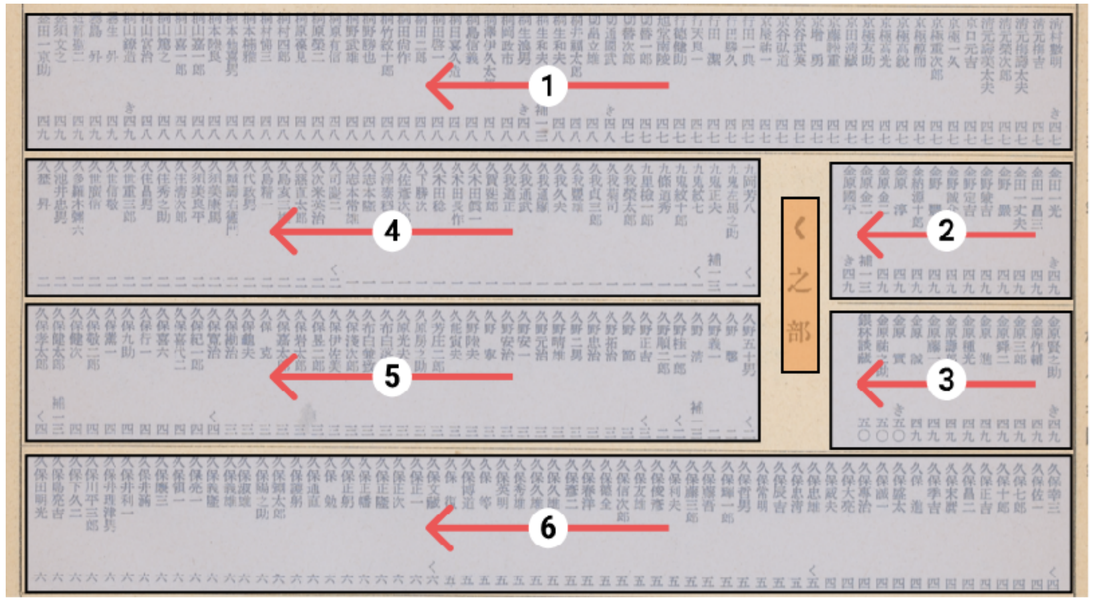


The four stages in layout element annotation. Our method detects the coordinates of the page frames, row regions, and text blocks. A text block classifier is then used to predict the block categories (indicated by the different colors in the figure), and the detections are refined accordingly. Reading orders and hierarchical dependency are generated for all layout elements. Finally, human annotators check the results and correct the errors.

Reading Order Generation



▲ Examples of the layout annotations and their reading order.



▲ Irregular reading orders in the index pages. The section header in row 2 and 3 disrupts the reading order

Pre-Trained Models

- We provide weights fro three models pre-trained on the HJDataset, namely, Faster R-CNN, Mask R-CNN, and RetinaNet.
- The implementation of the models is based on Detectron2, and the training details are stored in the corresponding configuration files for reproduction purposes.
- The table on the right shows the model performance (Average Precision) for different classes of objects in the dataset.

Category	Faster R-CNN	Mask R-CNN[a]	RetinaNet
Page Frame	99.046	99.097	99.038
Row	98.831	98.482	95.067
Title Region	87.571	89.483	69.593
Text Region	94.463	86.798	89.531
Title	65.908	71.517	72.566
Subtitle	84.093	84.174	85.865
Other	44.023	39.849	14.371
mAP	81.991	81.343	75.223

▲ Detection mAP @ IOU [0.50:0.95] of different models for each category on the test set. All values are given as percentages.

Thank you!

And don’t forget to check our project website.