The input data looks like this:

Iterate over the input folder recursively. Look for subfolders that contain identically named JSONs and JPGs. There is one of these for every page of the book that we are processing.

In a folder there are identically named JPG and JSON files for every page of the data. Within the JSON there is an array called “shapes”. Within the “shapes” each element corresponds to a part of the table; it contains either a single number, or a row of numbers (numbers or dashes representing there is no number in the row). The bounding boxes are stored in “points”. Consider only those shape elements that have a field called “human\_output”. Within this the field “human\_corrected\_text" contains the clean content of the cell.

I attach a file which shows how the input data looks like.

I would like you to write a script that iterates over the input folder. For every “human\_corrected\_text” element cut out the content of the image from the JPG. From these snippets generate the input data for training EffOCR.

The training data that you generate does not have to preserve the input folder structure. It is much better if you just dump all the snippets together into the same folder.