Improvements to direct pdf extraction program

The point of this program:

Some papers don't appear to have image websites, so in these special cases the images will have to be taken directly from the pdfs and stored somewhere. All attempts so far have used the fitz library

The current attempt is:

Find pages with pattern required. For each pattern found on the page find its dimensions. From the dimensions roughly guess the dimensions of each patterns image above it and download (using the pattern to define the filename).

This is what is currently implemented, the rough guess is the problem with this implementation. There are also page.get_images() functions that can be used but at the time it seemed easier as get_images() doesn't pick up vector images.

Improved solution (hasn't been implemented yet):

Find all caption locations in each page. Run get_images() and get_drawings() for the whole page. Then for each region bounded by caption locations use the reduce() function to get a union of all vectors images in each bounded region. The images and drawings in each region can then be associated with the captions bellow them. They can then be downloaded and named in appropriate ways.

This approach assumes that the document only has one column. Its much better than the current one as it uses the more efficient and effective get_images() when it can, the cropping will be much more accurate and cases like in the image bellow will be evaluated properly.

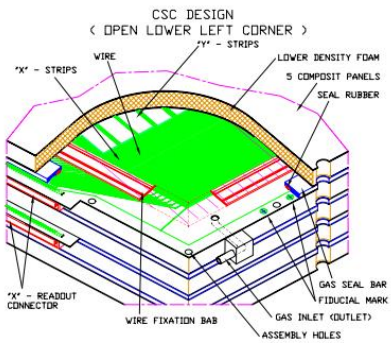

Figure 6.19: Structure of the CSC.

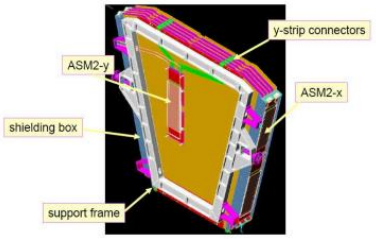

Figure 6.20: Model of a CSC chamber with four planes showing the location of the readout electronics.