Please refer to the attached permits for reference while going through the instructions.

The objective of this project is to automate the extraction of permit requirements/conditions from the PDF document of a permit into a structured format using machine learning models. To train the models, we need the annotations corresponding to every piece of information that needs to be extracted from the document.

The task consists of 2 parts –

1. Identify if the page contains any permit requirements/conditions or not. If not, leave the page empty. Some documents may contain supplementary permit-related information like applications, memorandum, forms, letters, attachments, etc. which are of no interest to us and need to be ignored while annotating the document. Example – for *AQ0105GPA06P.pdf*, pages 6-43 need to be annotated and for *142x-233-0030-S-01.pdf*, pages 3-9 need to be annotated.

Some pages might contain other requirements/conditions (such as qualifying requirements). Ignore such requirements/conditions. Example – leave page 5 empty for *AQ0105GPA06P.pdf*.

1. If the page contains any requirements/conditions, draw bounding boxes for the following categories of requirement/condition text in each image of the document (see examples of each below) –
   1. Number – The number of the requirement/condition if present.
   2. TextLevel(N) – Block of text with its corresponding level (N). The main requirement/condition corresponds to level 1. Each item directly under it corresponds to level 2 and so on. There are no limitations to the depth of the levels. A change in level can be identified by indentation, font weight, underlining, ordered/unordered listing, etc.
   3. ContinuousBlock – Block of text that is at the same level as the previous block and is also a part of it.

**Good Examples**

1. A screenshot of a computer

   Description automatically generated
2. A screenshot of a computer

   Description automatically generated with medium confidence
3. A screenshot of a computer

   Description automatically generated with medium confidence
4. A screenshot of a computer

   Description automatically generated with medium confidence
5. A screenshot of a computer

   Description automatically generated with medium confidence
6. A screenshot of a computer

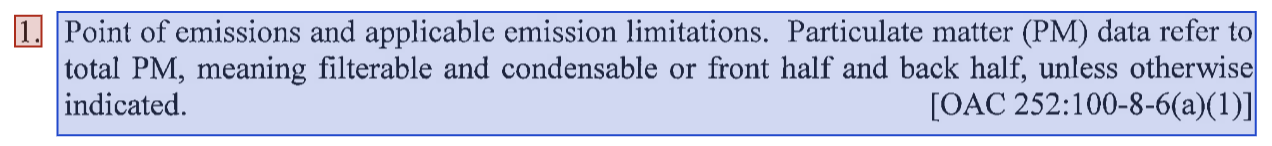
   Description automatically generated with medium confidence

Important notes –

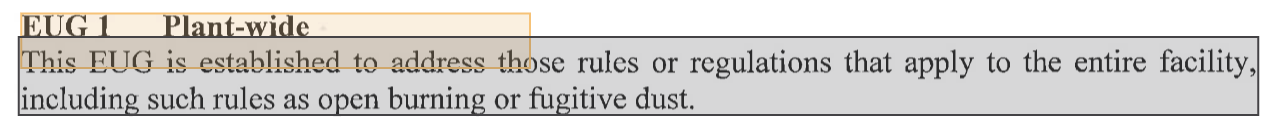
* Examples 1 shows all categories of labels correctly annotated.
* Certain requirements may contain references (in square brackets) below the text blocks. Include them in the bounding boxes. Examples 1 & 3.
* Subsequent header blocks at the same level are considered continuous. See annotations for ContinuousBlock in Examples 1 & 3.
* The levels are continuous across a document and therefore, it is important to annotate the images in the sequence in which they appear in the PDF document. All images for a document will appear in sequence during annotation and it is important to be mindful of it and refer to neighboring images while annotating an image. For example, Examples 1 and 2 are images from the same document in order. While trying to identify the level of the first text block in Example 2, the annotator needs to see that the last text block in the previous image (in Example 1) is at level 3, and since the text block in Example 2 is at a lower level, its level is 4.
* Examples 5 & 6 show non-numbered/bulleted requirements/conditions.
* A change in the level does not necessarily mean a change in indentation. Annotators need to look for all types of indicators. Example 4 shows a change in the level through font weight and the presence of a number in the header. Example 6 shows a change in the level through underlining of the text.
* Ignore any figures, tables, table descriptions or notes, and texts before the first requirement/condition. Example 4.

**Bad Examples**

* Do not exclude numbers/indexes from the bounding box of the text block.



* Make sure the bounding boxes are precise. Do not leave additional space or cut out pieces of texts from the boxes.

****

* Do not combine multiple blocks into one.

**A picture containing text, font, screenshot

Description automatically generated**