Table Extraction

This competition consists of the task of extracting structured data (tables) from unstructured documents (PDF files).

The competition is split up into two sub-competitions, **table detection** and **table structure recognition**. Entrants may choose to enter either one, or both.

The dataset includes a total of 150 tables: 75 tables in 27 excerpts from the EU and 75 tables in 40 excerpts from the US Government.

You will find the following folders:

- PDF – this folder contains the original PDF files from which the tables need to be extracted.
- GroundTruth – this folder contains 2 XML files for each PDF file, 1 for each sub-competition. They contain the true position and structure of every table in the PDFs. They should be used for training and evaluation.
- JSON – this folder contains JSON files with the absolute position of words on each page of each PDF file, as extracted by pdfminer. This is included for convenience and does not necessarily need to be used.
- PNG – this folder contains images of each page of each PDF file. This is included for convenience and does not necessarily need to be used.

Table detection sub-competition: Region model

In this sub-competition, entrants will be required to find the rectangular bounding-box coordinates of all tables in the dataset.

Table regions are defined as rectangular areas of a given page by their coordinates. Since a table can span more than one page, several regions can belong to the same table. For example, the ground truth file for a document with a table spanning from the first to the second page may look as follows:

```
<?xml version="1.0" encoding="UTF-8"?>
<document filename='filename.pdf'>
  <table id='0'>
    <region id='0' page='1'>
      <bounding-box x1='87' y1='117' x2='551' y2='220'/>
    </region>
    <region id='1' page='2'>
      <bounding-box x1='87' y1='261' x2='551' y2='364'/>
    </region>
  </table>
  <table id='1'>
    ...
```

```
  </table>
  ...
</document>
```

For each tabular region that is found, entrants are **only required to return its rectangular bounding-box in PDF coordinates.** Note that the page-numbering is 1-based (all document excerpts in the example begin with page 1). The region and table ID numbering is 0-based; tables within a document and regions within a table can be output in any order.

Table structure recognition sub-competition: Cell structure model

The aim of the table structure recognition sub-competition is to determine the cell structure of tables *given correct information about their location*. This means that you can use the ground truth about table location when solving this problem. It is therefore permissible to participate in this sub-competition without solving the table location sub-competition. Even if you have solved the table detection task, we recommend entrants to use the ground truth information regarding the table locations when solving this one, in order to avoid unnecessarily detection errors.

The cell structure of a table is defined as a matrix of cells. Cells are defined by their textual content and their start and end column and row positions. Blank cells are not represented in this format. An example looks like this:

```
<?xml version="1.0" encoding="UTF-8"?>
<document filename='filename.pdf'>
  <table id='0'>
    <region id='0' page='3' col-increment='0' row-increment='0'>
    <cell id='0' start-row='0' start-col='0'>
      <bounding-box x1='70' y1='79' x2='131' y2='91'/>
      <content>COUNTRY</content>
    </cell>
    <cell id='1' start-row='0' start-col='1' end-col='2'>
      <bounding-box x1='165' y1='79' x2='201' y2='91'/>
      <content>3 years</content>
    </cell>
    <cell id='2' start-row='0' start-col='3'>
      <bounding-box x1='234' y1='79' x2='271' y2='91'/>
      <content>4 years</content>
    </cell>
    ...
    </region>
    ...
  </table>
  ...
</document>
```

In the ground truth for the example dataset, the table numbers correspond to those in the relevant region model files. In this competition, this need not be the case, as entries for the table structure recognition sub-competition will be evaluated independently of the table location competition.

In contrast to the region model, for the cell structure model, entrants are required to return the **textual content** (<content> tag) for each cell; the <bounding-box> tag can be ignored.

The cell numbering begins at (0,0) for the top-left cell. The attributes end-col and end-row are optional; if they are omitted, the col and/or rowspan are assumed to be 1.