# [How to extract tables from pdf file](http://answers.opencv.org/question/63847/how-to-extract-tables-from-an-image/)

**Preface**

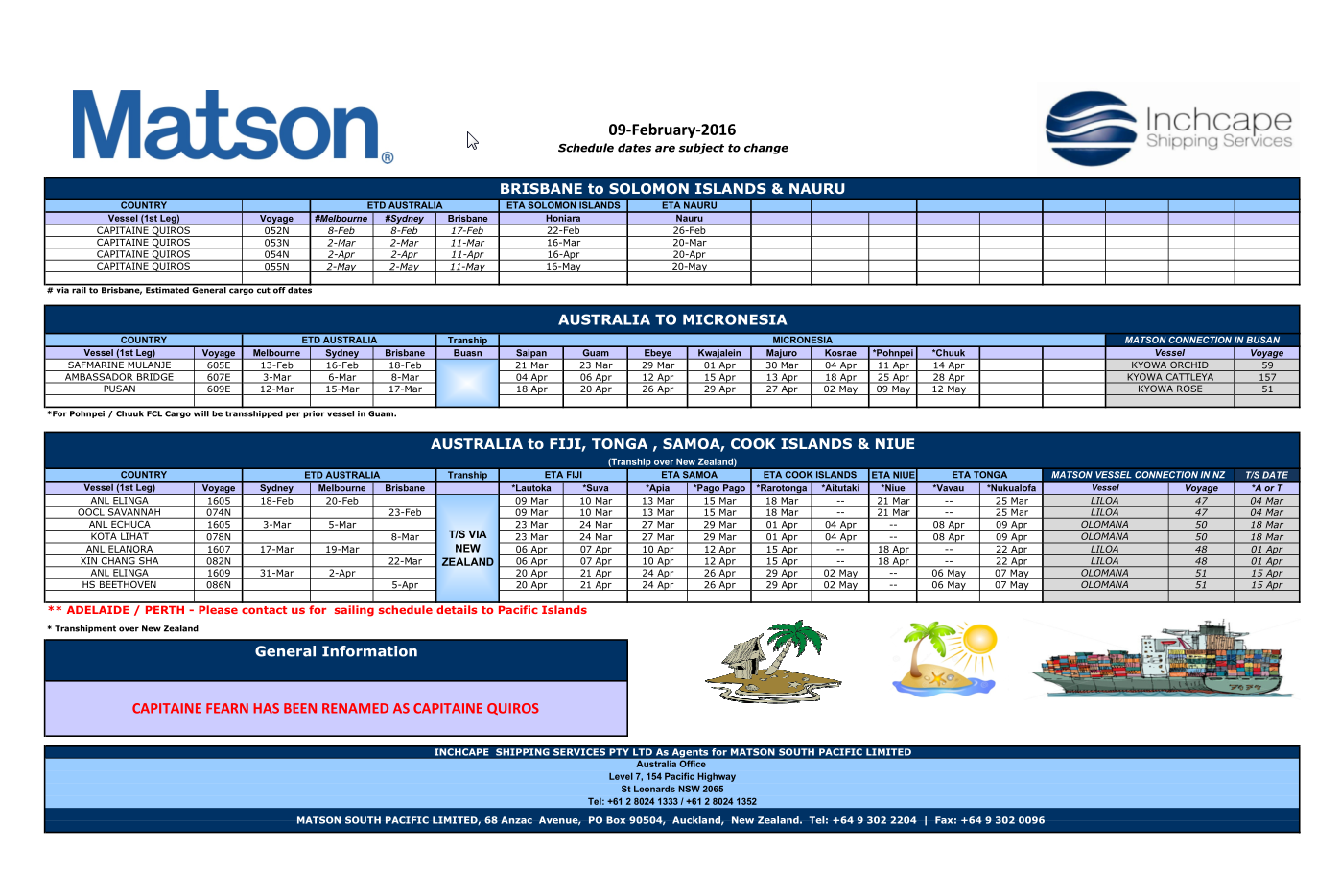
Due to business needs, we met the demand of changing all the PDF to HTML.

Realize that there are few open source Java API to do this, while the rest of the API can only gain character, not the location of the access form.

We are looking for ways to get the results. And hope it can adapt to the situation as much as possible

**Face the situation**

We met the PDF form example as below:



We can see that the form has obvious horizontal and vertical line. So I think, if we can get the horizontal and vertical line, then we can know their intersection, for each grid location. So we can use it to get location information

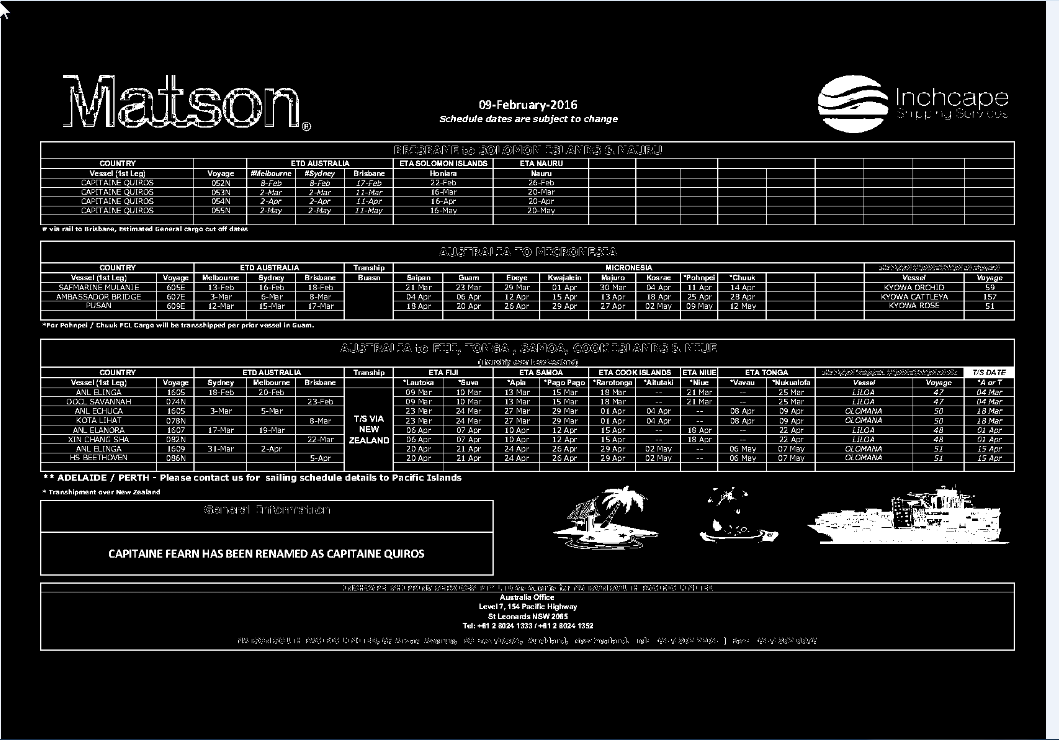
**The realization of the algorithm**

1. extract images from pdf

First of all, we use “pdfbox” to extract every page to PNG format images. Into the cause of the picture is in order to get the location information according to the pixel

1. Transform source image to gray if it is not

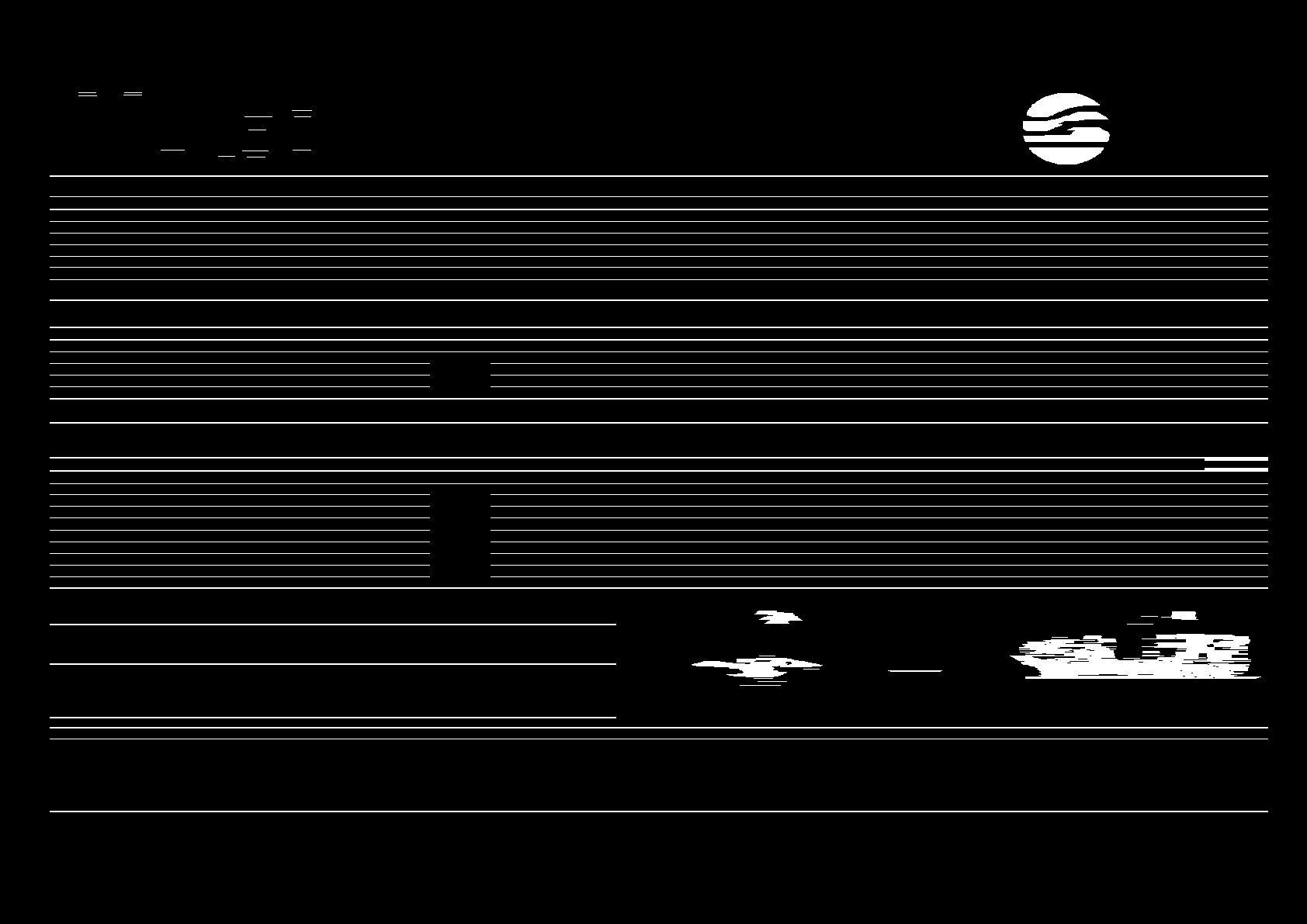
The reason for this is that there are only two RGB, better able to discern the need of the pixel.

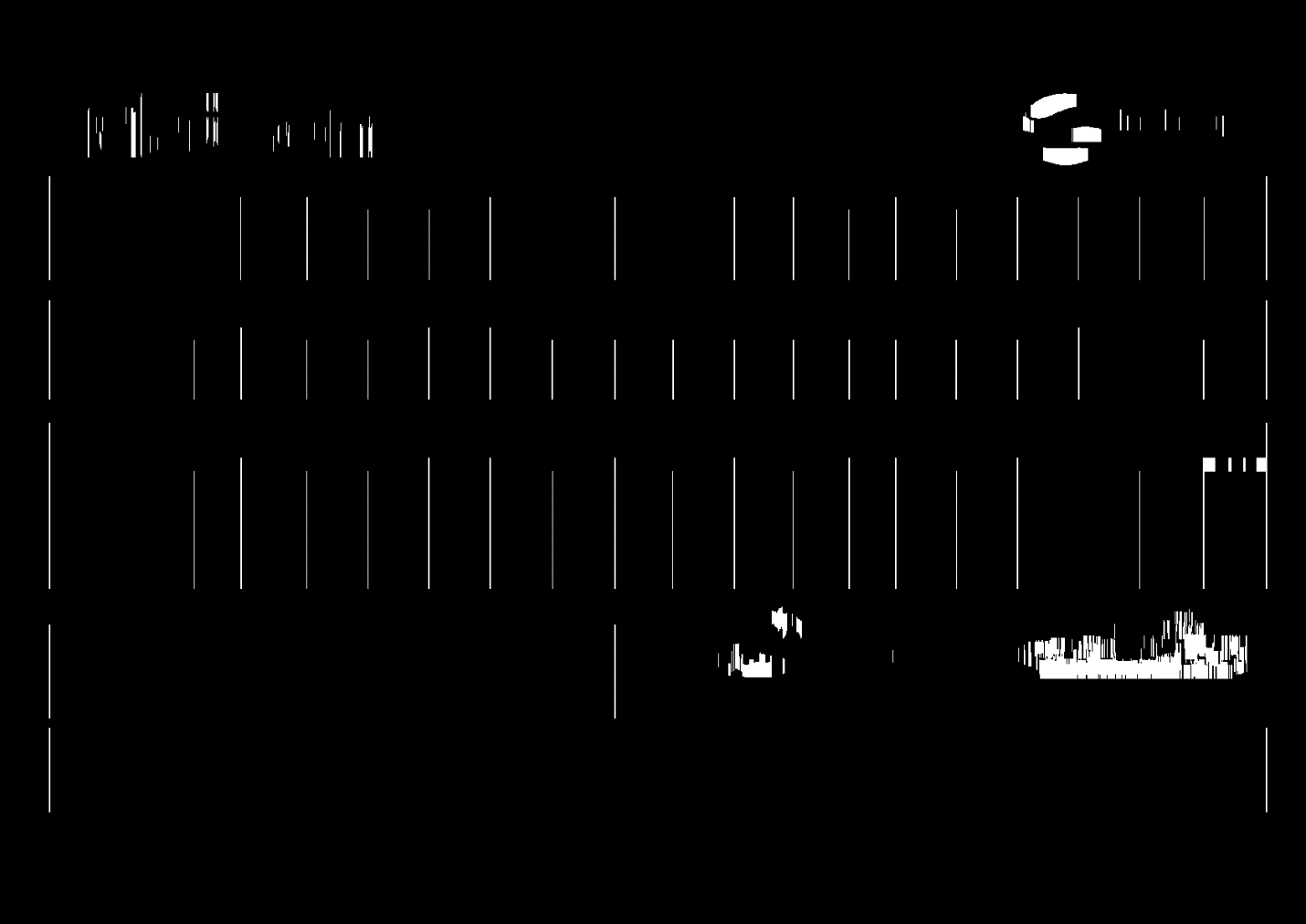


1. extract the horizontal and vertical lines

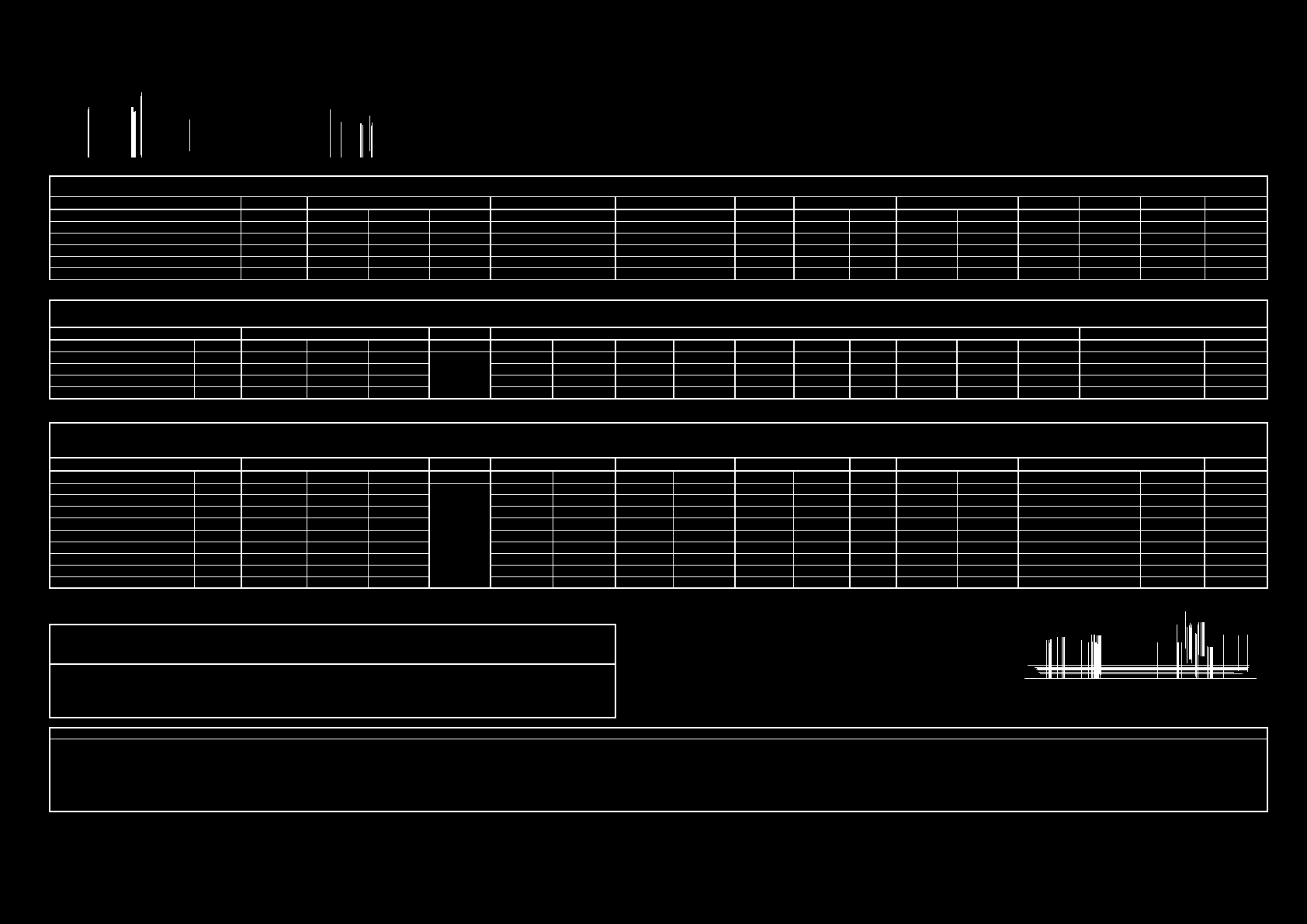
We can set a scale, continuous pixels can be thought of as a horizontal line or vertical line

So we can get the result like this.



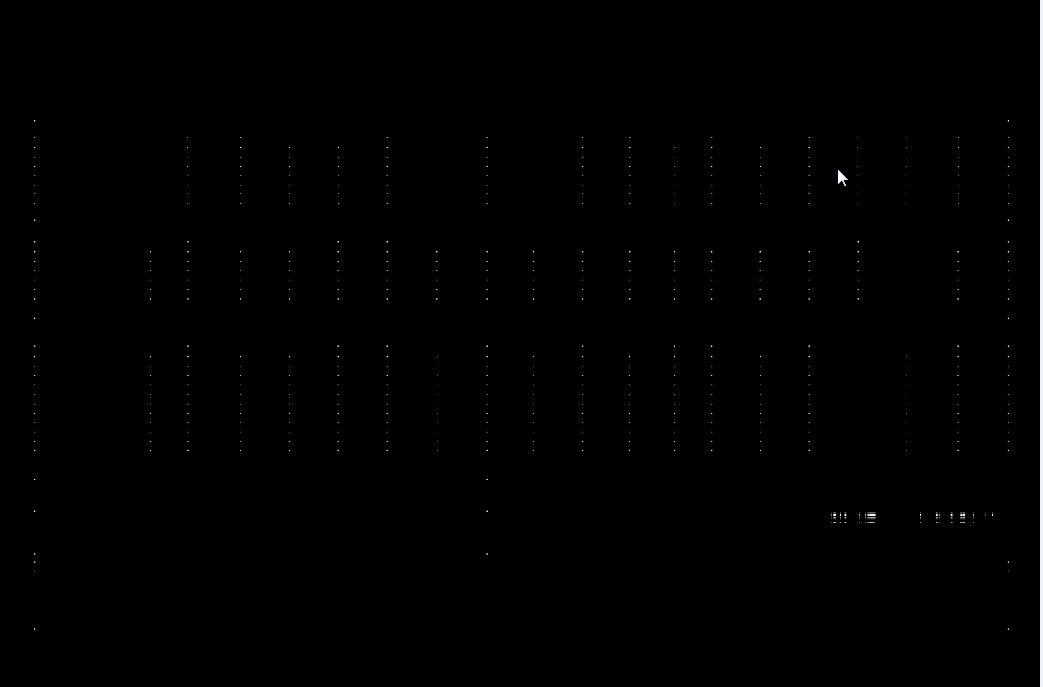


From the horizontal and vertical line intersection, we can get form. And split different forms base on first vertical pixel.



4. Obtain the intersection matrix

Of course, our goal is not to form, but all of the intersection. Find the joints between the lines of the tables, we will use this information in order to distinguish tables from pictures (tables will contain more than 4 joints while a picture only 4



After winning matrix, through filtering algorithm, turn it into single pixel

5. Base on single points, we can get areas table. And then the HTML form

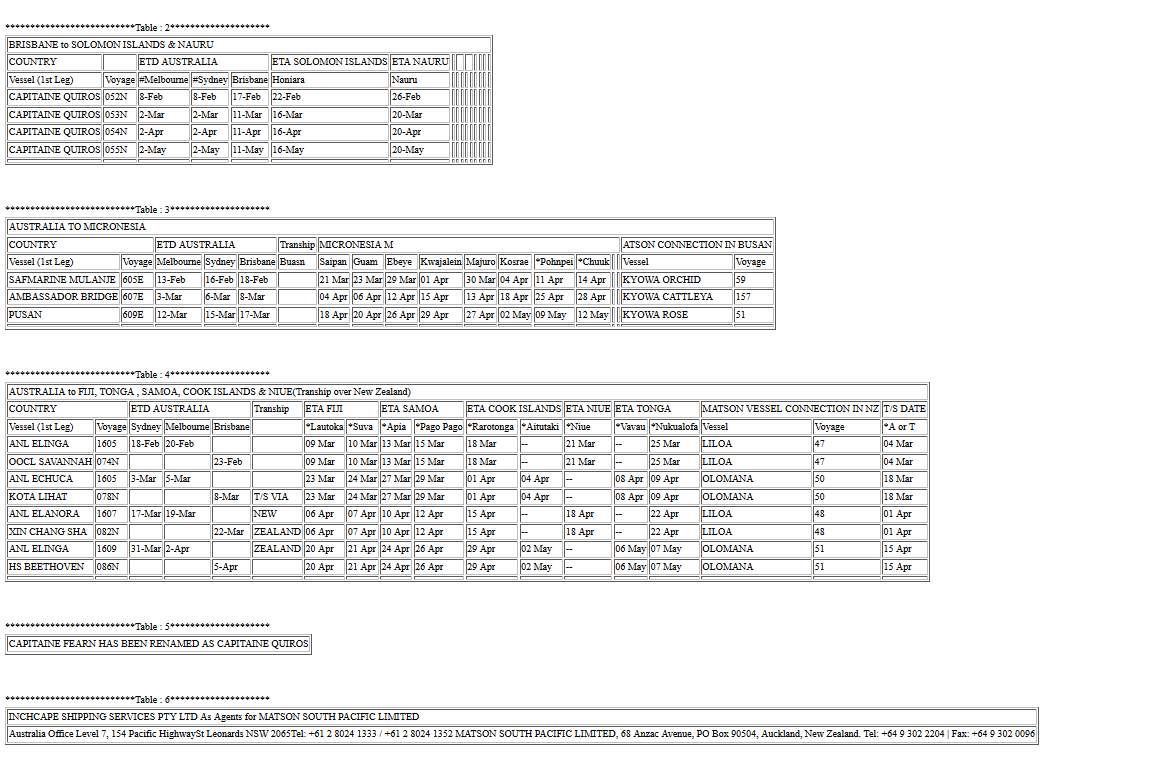
**Specific processing methods:**

1. Each point into a single standard (x , y )coordinates.
2. Use Y coordinates as order, put the coordinates of each row in the array, each number of the adjacent two arrays of contrast. According to the size of small array as a standard for a lattice of four points.
3. To obtain the position information to the form, use “StringBuffer” joining together into HTML form, containing the location information of the form.

**Ps**: For an example: <td> x coordinates, y coordinates, td width, td height<td>

1. Because of the characters in the PDF position information and pictures of the location information is proportional relationship. So we can according to the location information of the form, read the PDF file location corresponding character content, form and replace the position of the information. This result is that we need to complete forms. PDF provides according to the coordinates, and access to the designated area method of the characters.

6. The results of example

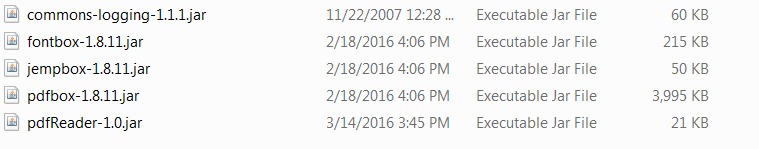


**Ps**: There are some information is not necessary, we can simple filter it

# How to use this program

## **Import the jar package**

(1)We put the program into a callable jars, so you can use it by Import the jar package below



**The sources :**

(2)If you project is maven project, you can just import pdfReader-1.0.jar

And set pom.xml like:

<dependency>

<groupId>org.apache.pdfbox</groupId>

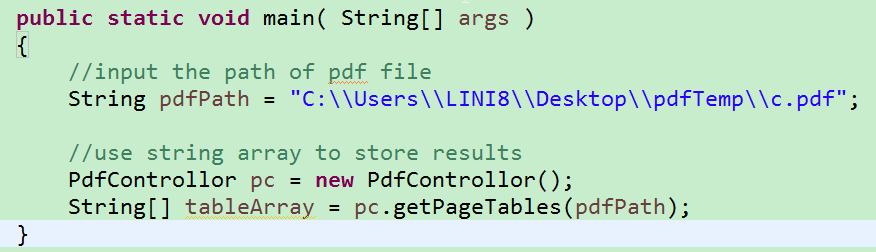
<artifactId>pdfbox</artifactId>

<version>1.8.11</version>

<dependency>

## **Code calls**

There is an example:



## **about config file “PdfImgConfig.properties”**

we use “PdfImgConfig.properties” to set properties , when first you run The program, Program will generate a file named “PdfImgConfig.properties” under the root directory of the project, you may need to refresh the project to see it.

## **The configuration file of the tutorial ,“PdfImgConfig.properties”**

1. delete file or not, after run pdf. use TRUE OR FALSE

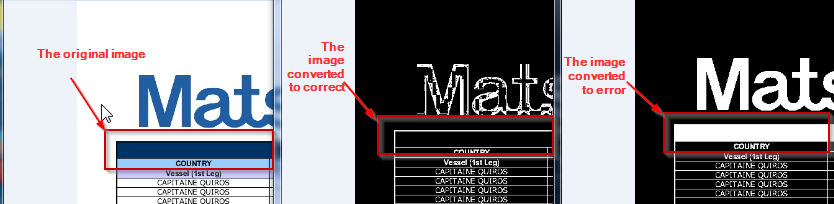
DELETE\_PDF = FALSE

DELETE\_PNG = FALSE

DELETE\_FOLDER = FALSE

1. ignore color format rgb , when one table missing some line , perhaps there are some color can't correct converted to black and white

* RGBIGNORE = 0,51,101 & 0,50,101 & 33,94,161 & 0,111,192 & 0,112,192
* For example



If the images can converted correct, we will not get the result.just like the form above. So we should ignore blur color (rgb = 0,51,101).

* Ps: how to get rgb.

1. Find images named (e.g 0.png), open png with firefox
2. Use firefox “colorpicker” plug-in to get the rgb you want to ignore
3. the boundary of rgb, split to 2 different color base this RGB

R = 150

G = 150

B = 150

1. the length of min vertical size. Means the vertical Less than the minimum pixel length, do not think that is the line

VERTICAL\_LENGTH = 40

1. the length of min horizontal size, Means the horizontal Less than the minimum pixel length, do not think that is the line

HORIZONTAL\_LENGTH = 250

1. the scale to recognition different vertical line (Two pixels vertical distance)

VERTICAL\_SPLIT\_TABLE\_SCALE = 5

1. the scale to remove pixels if is too close , in order every points are single pixel

REMOVE\_PIXELS\_STANDARD= 3

1. if pixels around standard pixel,change value to standard,this is the around scale

CHOOSE\_PIXELS\_STANDARD= 10