1. XML File Structure

Below is an example XML file

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE document SYSTEM "dtd.dtd">

<document>
    <document_summary no_pages="1"/>
    <page page_id="1" image_filename="mp00088bw.tif">
        <page_summary no_text_regions="22" no_image_regions="0"
            no_line_drawing_regions="0" no_graphic_regions="0"
            no_table_regions="0" no_chart_regions="0"
            no_separator_regions="0" no_maths_regions="0"
            no_frame_regions="0" no_noise_regions="0"/>
        <page_pixel_size width="2340" height="3135"/>
        <text_region id="1" txt_orientation="0"
            txt_reading_direction="Left_To_Right"
            txt_leading="" txt_kerning=""
            txt_font_size="12" txt_type="Paragraph" txt_colour="Black"
            txt_reverse_video="No" txt_indented="No"
            txt_primary_lang="English" txt_secondary_lang="None"
            txt_primary_script="Latin" txt_secondary_script="None"
            txt_bgcolour="White" txt_reading_orientation="0">
            <coords no_coords="4">
                <point x="10" y="10"/>
                <point x="20" y="10"/>
                <point x="20" y="20"/>
                <point x="10" y="20"/>
            </coords>
        </text_region>
    </page>
</document>
```

The main entity here is a *Document*, which is the only type of entity that can be found in an XML file after the header lines. Inside the Document (between the `<document>` and `</document>` tags) two types of entities are allowed: the *Document Summary* and a number of *Pages*.

1.1 Document Summary Tags

The document summary section specifies how many pages there are in the document.

```
<document_summary no_pages="1">
```

no_pages
The no_pages attribute specifies how many pages the XML file represents. Currently, this should be set to 1.

1.2 Page Attributes & Child Elements

The most important parts of every document are the pages it contains. Each page is represented here as a separate entity, and information about each page is given between the `<page>` and `</page>` tags. Currently, only 1 page per document is supported.

Page ID Attribute
The page ID uniquely identifies the page within the document

```
<page id="1"></page>
```

Each page can be decomposed into a number of regions. The regions defined cover a wide range of document content and are presented in the next section

Image Filename Attribute
Image filename attributes are used to indicate the name of the file used in the segmentation. The path to the file should not be included, as this may vary from person to person.

```
<page image_filename="ta000012.tif"></page>
```

Page Size
The page size tags define the width and height in pixels of the page.

```
<page_pixel_size width="2340" height="3135"/>
```

1.3 Page Summary Tags

The Page Summary contains the number of occurrences of each type of region in the page. Such information can be valuable when searching for pages containing certain types of regions, since only the page summary has to be accessed.

```
<page_summary no_text_regions="22" no_image_regions="0"
no_line_drawing_regions="0" no_graphic_regions="0" no_table_regions="0"
no_chart_regions="0" no_separator_regions="0" no_maths_regions="0"
no_frame_regions="0" no_noise_regions="0"/>
```

no_text_regions
The no_text_regions attribute specifies how many regions of text the page element contains. Such a region represents any part of the image which is considered to represent text, such as image captions, drop capital, header/footers and headings.

no_image_regions
The no_image_regions attribute specifies how many images the page element contains.

no_line_drawing_regions
The no_line_drawing_regions attribute specifies how many line drawings the page element contains.

no_graphic_regions
The no_graphic_regions attribute specifies how many graphics regions the page element contains. Such a region represents any part of the image which is considered to represent a graphic.

no_table_regions
The no_table_regions attribute specifies how many tables the page element contains. These regions contain tabular data, either with or without a surrounding frame.

no_chart_regions
The no_chart_regions attribute specifies how many charts the page element contains. Although appearing to be a type of graphics, charts are labelled separately as they can contain suplemental information to the document which can be extracted.

no_separator_regions
The no_separator_regions attribute specifies how many separators the page element contains. Separators are lines or equivalent divisions which can run between columns of text.

no_maths_regions
The no_maths_regions attribute specifies how many equations the page element contains. An equation region should contain the entire equation, grouping together each individual part.

no_frame_regions
The no_frame_regions attribute specifies how many frame regions the page element contains.

no_noise_regions
The no_noise_regions attribute specifies how many regions of noise the page element contains. Noise is any artifact which the scanner has introduced, or existed on the original page, which is not part of the document. All noise should be labelled, though this can prove time consuming.

2. Region Types

All regions contain a unique ID number to identify them in the document, and also all contain coordinate sets to define the outline of the region. The coordinate list of each region is the only entity that must appear between the opening and closing tags of the region, apart from frame regions which contain sub-regions as well as coordinate sets.
It is required that each region has an ID attribute, but it is not necessary that each region has each attribute supplied.

```
<coords no_coords="4">
    <point x="10" y="10"/>
    <point x="20" y="10"/>
    <point x="20" y="20"/>
    <point x="10" y="20"/>
</coords>
```

The only attribute of the coordinate set is “no_coords”, which is the number of points given. The actual pairs of x, y values appear between the `<coords>` and `</coords>` tags, each in a separate "point" element which has just two attributes – "x" and "y".

2.1 Chart region

If the region surrounds a part of the document, which contains a chart or graph of some type, then the region type is to be set to “chart”.

In the examples given below, only the opening tag (with the list of attributes) is given for each region. In the actual file, this is followed by a coordinate list, as described earlier, and the appropriate closing tag.

```
<chart_region id="1" chart_emb_text="Yes" chart_orientation="0"
chart_no_colours="2" chart_type="Pie" chart_bgcolour="White">
```

ID

Meaning : The unique id number of this region.
Type : Integer
Default : No Default (Increases Sequentially)
Required : Yes

Embedded Text

Meaning : Specifies whether the chart region also contains text.
Type : Boolean
Default : No Default
Required : No

Chart Orientation

Meaning : Specifies the orientation of the baseline of the rectangle that encapsulates the chart region.

Type : Floating Point
Default : No Default
Required : No
Range : [+ 90, - 89]
Units : Degrees

Number of colours

Meaning : Specifies an approximation of the number of colours used in the chart region.
Type : Integer
Default : No Default
Required : No

Chart Type

Meaning : Specifies the type of chart used in the chart region.
Type : List (See Chart Type List)
Default : No Default
Required : No

Background colour

Meaning : Specifies an approximation of the background colour of the chart region.
Type : List (See colour approximation list)
Default : White
Required : No

2.2 Frame region

If the region surrounds a side bar or some other part of document which is enclosed in a frame and is to be considered as separate from the rest of the document. Frame regions are intended to wrap around other regions that constitute the frame.

```
<frame_region id="2">
```

Currently, a frame region only supports a region id attribute.

2.3 Graphic region

A graphic is considered to be a simple graphic, such as a company logo or illustrated text.

```
<graphic_region id="3" gfx_type="Other" gfx_emb_text="No"
gfx_orientation="0" gfx_no_colours="2">
```

ID

Meaning : The unique id number of this region.
Type : Integer
Default : No Default (Increases Sequentially)
Required : Yes

Type

Meaning : Specifies the type of graphic in the region.
Type : List (See Graphic Region Type List)
Default : No Default
Required : No

Embedded Text

Meaning : Specifies whether the graphic region also contains text.
Type : Boolean
Default : No Default
Required : No

Graphic Orientation

Meaning : Specifies the orientation of the baseline of the rectangle that encapsulates the graphic region.
Type : Floating Point
Default : No Default
Required : No
Range : [+ 90, - 89]
Units : Degrees

Number of colours

Meaning : Specifies an approximation of the number of colours used in the graphic region including the background colour.
Type : Integer
Default : No Default
Required : No

2.4 Image region

An image is considered to be more intricate and complex than a simple graphic. These can be photos or drawings and can be in millions of colours down to pure black and white.

```
<image_region id="4" img_colour_type="Black_And_White"
img_orientation="0" img_emb_text="No" img_bgcolour="White">
```

ID

Meaning : The unique id number of this region.
Type : Integer
Default : No Default (Increases Sequentially)
Required : Yes

Colour Type

Meaning : Specifies the depth/number of colours used in the image.
Type : List (See Colour Type List)
Default : No Default
Required : No

Image Orientation

Meaning : The orientation of the base line of the rectangle that encapsulates the image.
Type : Floating Point
Default : No Default
Required : No
Range : [+ 90, - 89]
Units : Degrees

Embedded text
Meaning : Specifies whether the image region also contains text.
Type : Boolean
Default : No Default
Required : No

Background colour
Meaning : Specifies an approximation of the background colour of the image region.
Type : List (See colour approximation list)
Default : White
Required : No

2.5 Line drawing region

A line drawing is an illustration in black and white without solid areas. These can be items such as diagrams

```
<line_drawing_region id="5" drwg_emb_text="No" drwg_orientation="0"
drwg_pen_colour="Black" drwg_bgcolour="White">
```

ID
Meaning : The unique id number of this region.
Type : Integer
Default : No Default (increases sequentially)
Required : Yes

Embedded text
Meaning : Specifies whether the line drawing region also contains text.
Type : Boolean
Default : No
Required : No

Drawing Orientation
Meaning : Specifies the orientation of the baseline of the rectangle that encapsulates the drawing region.
Type : Floating Point
Default : No Default
Required : No
Range : [+ 90, - 89]
Units : Degrees

Pen colour
Meaning : Specifies an approximation of the colour of the pen used to create the line drawing.

Type : List (See colour approximation list)
Default : Black
Required : No

Background Colour

Meaning : Specifies an approximation of the background colour of the drawing region.
Type : List (See colour approximation list)
Default : White
Required : No

2.6 Maths region

Although basically textual, areas containing equations and mathematical symbols are treated slightly differently and are to be labeled as maths regions

```
<maths_region id="6" maths_bgcolour="White" maths_orientation="0">
```

ID

Meaning : The unique id number of this region.
Type : Integer
Default : No Default (Increases Sequentially)
Required : Yes

Background Colour

Meaning : Specifies an approximation of the background colour of the maths region.
Type : List (See colour approximation list)
Default : White
Required : No

Orientation

Meaning : Specifies the orientation of the baseline of the rectangle that encapsulates the maths region.
Type : Floating Point
Default : No Default
Required : No
Range : [+ 90, - 89]
Units : Degrees

2.7 Noise region

A noise region denotes an area where no real data lies, only false data created by artifacts on the document or scanner noise. A noise region does not have any properties other than the region id.

```
<noise_region id="7">
```

ID

Meaning : The unique id number of this region.
Type : Integer
Default : No Default (Increases Sequentially)
Required : Yes

2.8 Separator Region

Separators are lines that lie between columns and paragraphs and can be used to logically separate different articles from each other.

```
<separator_region id="8" sep_orientation="0" sep_colour="Black"
sep_bgcolour="White">
```

ID

Meaning : The unique id number of this region.
Type : Integer
Default : No Default (Increases Sequentially)
Required : Yes

Separator Orientation

Meaning : Specifies the orientation of the separator contained in the region.
Type : Integer
Default : No Default
Required : No
Range : [+ 90, - 89]
Units : Degrees

Separator colour

Meaning : Specifies an approximation of the colour of the separator in the separator region.
Type : List (See colour approximation list)
Default : Black
Required : No

Background Colour

Meaning : Specifies an approximation of the background colour of the separator region.
Type : List (See colour approximation list)
Default : White
Required : No

2.9 Table Region

Tabular data in any form is represented with a table region. Rows and columns may or may not have separator lines. These lines are not separator regions however.

```
<table_region id="9" tbl_rows="" tbl_columns="" tbl_line_colour="Black"
tbl_orientation="0" tbl_line_separators="Yes" tbl_bgcolour="White"
tbl_emb_text="Yes">
```

ID

Meaning : The unique id number of this region.
Type : Integer
Default : No Default (Increases Sequentially)
Required : yes

Rows

Meaning : Specifies the number of rows present in the table.
Type : Integer
Default : No Default
Required : No

Columns

Meaning : Specifies the number of columns present in the table.
Type : Integer
Default : No Default
Required : No

Line colour

Meaning : Specifies the colour of the lines used in the table.
Type : List (See colour approximation list)
Default : Black
Required : No

Table orientation

Meaning : Specifies the orientation of the base line of the table region.
Type : Floating Point
Default : No Default
Required : No
Range : [+ 90, - 89]
Units : Degrees

Line separators

Meaning : Specifies the presence of line separators in the table.
Type : Boolean
Default : No Default
Required : No

Table Background Colour

Meaning : Specifies the background colour of the table
Type : List (See colour approximation list)
Default : White
Required : No

Embedded Text

Meaning : Specifies whether the table region also contains text.
Type : Boolean
Default : No Default
Required : No

2.10 Text Region

Pure text is represented as a text region. This includes drop caps, but particularly ornate text may be considered as a graphic.

```
<text_region id="10" txt_orientation="0" txt_reading_orientation="0"
txt_reading_direction="Left_To_Right" txt_leading="" txt_kerning=""
txt_font_size="12" txt_text_type="Paragraph" txt_text_colour="Black"
txt_reverse_video="No" txt_indented="No" txt_primary_language="English"
txt_secondary_language="None" txt_primary_script="Latin"
txt_secondary_script="None" txt_bgcolour="White">
```

ID

Meaning : The unique id number of this region.
Type : Integer
Default : NA (increases sequentially)
Required : Yes

Orientation

Meaning : Specifies the orientation of a straight-line segment passing through all text segments.
Type : Floating Point
Default : No Default
Required : No
Range : [+ 90, - 89]
Units : Degrees

Reading Orientation

Meaning : The degrees by which you need to turn your head in order to read the document when it is placed on the horizontal.
Type : Floating Point
Default : No Default
Required : No
Range : [0, 180]
Units : Degrees

Reading Direction

Meaning : Specifies the direction in which text in the text region should be read.
Type : List (See Reading Direction List)
Default : Left_To_Right
Required : No

Leading

Meaning : The degree of space between lines of text.
Type : Integer
Default : No Default
Required : No
Units : Points

Kerning

Meaning : The degree of space between the characters in a string of text.
Type : Integer

Default : No Default
Required : No
Units : Points

Font size

Meaning : The size of characters used in a string of text.
Type : Integer
Default : No Default
Required : No
Units : Points

Text type

Meaning : Defines the nature of text captured in a particular text region.
Type : List (See Text Type List)
Default : No Default
Required : No

Text colour

Meaning : Defines an approximation of the text colour captured in the region.
Type : List (See colour approximation list)
Default : Black
Required : No

Reverse Video

Meaning : When the colour of text appears reversed against a background colour.
Type : Boolean
Default : No
Required : No

Indented

Meaning : Defines whether a region of text is indented or not.
Type : Boolean
Default : No
Required : No

Primary Language

Meaning : Defines the primary language used in a text region.
Type : List (See language list)
Default : English
Required : No

Secondary Language

Meaning : Defines the secondary language used in a text region.
Type : List (See language list)
Default : No Default
Required : No

Primary script

Meaning : Defines the primary language script used in a text region.
Type : List (See script list)
Default : Latin
Required : No

Secondary script

Meaning : Defines the primary language script used in a text region..
Type : List (See script list)
Default : No Default
Required : No

Background colour

Meaning : Specifies an approximation of the background colour of the text region.
Type : List (See colour approximation list)
Default : White
Required : No

3. Chart Type List:

- Pie
- Line
- Other

4. Graphic Region Type List:

- Logo
- Letterhead
- Handwritten_Annotation
- Stamp
- Signature
- Paper_Grow
- Punch_Hole
- Other

5. Colour Type List:

- Black_And_White
- 4_Bit_Greyscale
- 8_Bit_Greyscale
- 4_Bit_Colour
- 8_Bit_Colour
- 16_Bit_Colour
- 24_Bit_Colour
- 32_Bit_Colour

6. Reading Direction list

- Left_To_Right
- Right_To_Left
- Top_To_Bottom
- Bottom_To_Top

7. Text Type List:

- Paragraph
- Heading
- Sub_Heading
- Sentence
- Caption
- Header
- Footer
- Page
- Number
- Quote
- Drop_Capital

8. Colour approximation list:

- Black
- Red
- White
- Green
- Blue
- Yellow
- Orange
- Pink
- Grey
- Turquoise
- Indigo
- Violet
- Cyan
- Magenta

9. Language List & Script List:

Scripts shown in brackets.

- Afrikaans (latin)
- Albanian (latin)
- Amharic (Ethiopic)
- Arabic (Arabic)
- Basque (latin)
- Bengali (Bengali)
- Bulgarian (Cyrillic)
- Cambodian
- Cantonese (Traditional_Chinese)
- Chinese (Simplified_Chinese)
- Czech (latin)
- Danish (latin)
- Dutch (latin)
- English (latin)
- Estonian (latin)
- Finnish (latin)
- French (latin)
- German (latin)
- Greek (greek)
- Gujarati (Gujarati)
- Hebrew (Hebrew)
- Hindi (devangari)
- Hungarian (latin)
- Icelandic (latin)
- Indonesian (latin)
- Gaelic (latin)
- Italian (latin)
- Japanese (?)
- Korean (?)
- Latvian (latin)
- Malay (latin)
- Norwegian (latin)
- Polish (latin)
- Portuguese (latin)

- Russian (Cyrillic)
- Spanish (latin)
- Swedish (latin)
- Thai (thai)
- Turkish (latin)
- Urdu (Arabic)
- Punjabi (Gurmukhi)
- Welsh (latin)