A Short Introduction to the PRImA PAGE Format

Submissions of segmentation results for the ICDAR2009 Page Segmentation Competition should preferably be in the PRImA **PAGE** (**P**age **A**nalysis and **G**round truth **E**lements) format. The corresponding XML Schema can be found here:
http://schema.primaresearch.org/PAGE/gts/pagecontent/2009-03-16/pagecontent.xsd
In the following you will find a simple example for the XML format as well as further information on more specific elements. For a complete reference of elements please refer to the XML Schema definition (see above).

1. PAGE Example

```
<?xml version="1.0" encoding="UTF-8"?>
<pcGts mlns="http://schema.primaresearch.org/PAGE/gts/pagecontent/2009-03-16"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xsi:schemaLocation="http://schema.primaresearch.org/PAGE/gts/pagecontent/2009-03-16
      http://schema.primaresearch.org/PAGE/gts/pagecontent/2009-03-16/pagecontent.xsd">
  <pcMetadata>
    <pcCreator>PRImA Research Group</pcCreator>
    <pcCreated>2008-11-20T12:40:29</pcCreated>
    <pcLastChange>2009-03-16T12:30:15</pcLastChange>
  </pcMetadata>
  <page image_filename="mp00042c.tif" image_width="1367" image_height="2254">
           <text_region id="t1" txt_bgcolour="White"
                 txt_colour="Black" txt_font_size="12" txt_indented="No"
                 txt_kerning="" txt_leading="" txt_orientation="0"
                 txt_primary_language="English" txt_primary_script="Latin"
                 txt_reading_direction="Left_To_Right" txt_reading_orientation="0"
                 txt_reverse_video="No" txt_secondary_language="None"
                 txt_secondary_script="None" txt_type="Paragraph">
                 <coords>
                      <point x="10" y="10" />
                      <point x="20" y="10" />
                      <point x="20" y="20" />
                      <point x="10" y="20" />
                 </coords>
           </text_region>
           <noise_region id="n1">
                 <coords>
                      <point x="50" y="50" />
                      <point x="60" y="60" />
                      <point x="70" y="50" />
                      <point x="70" y="30" />
                      <point x="50" y="30" />
                 </coords>
           </noise_region>
           <text_region id="t2" txt_bgcolour="White"
                 txt_colour="Black" txt_font_size="12" txt_indented="No"
                 txt_kerning="" txt_leading="" txt_orientation="0"
                 txt_primary_language="English" txt_primary_script="Latin"
                 txt_reading_direction="Left_To_Right" txt_reading_orientation="0"
                 txt_reverse_video="No" txt_secondary_language="None"
                 txt_secondary_script="None" txt_type="Paragraph">
                 <coords>
                      <point x="100" y="150" />
                      <point x="200" y="150" />
                      <point x="200" y="200" />
                      <point x="100" y="200" />
                 </coords>
           </text_region>
  </page>
</pcGts>
```

2. General Structure

The root element is `<pcGts>` (page content Ground truth and storage) and must have the two child elements `<pcMetadata>` and `<page>`.

`<pcMetadata>` must contain `<pcCreator>` (information about the creator of this file), `<pcCreated>` (timestamp of the creation e.g. `2008-11-20T12:40:29`) and `<pcLastChanged>` (timestamp of the last change e.g. `2009-03-16T12:30:15`) and may also have an optional `<pcComments>` (any text) element.

`<page>` holds the actual layout data in form of region elements and some additional information about the organization of layout objects. Required attributes for `<page>` are `image_filename` (the file name of the image – without any path information – to which this segmentation result belongs to e.g. `mp00042c.tif`), image_width (width of the image e.g. `1367`) and image_height (height of the image e.g. `2254`).

The most important region elements as children of page are:

```
<text_region>
<image_region>
<line_drawing_region>
<graphic_region>
<table_region>
<chart_region>
<separator_region>
<maths_region>
<noise_region>
```

All of those region types may occur in any order and in any number (0..unbounded).

3. 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 possible attribute supplied.

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

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".

3.1 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

3.2 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

3.3 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

3.4 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

3.5 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

3.6 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

3.7 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

3.8 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

3.9 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

4. Chart Type List:

- Pie
- Line
- Other

5. Graphic Region Type List:

- Logo
- Letterhead
- Handwritten_Annotation
- Stamp
- Signature

- Paper_Grow
- Punch_Hole
- Other

6. 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

7. Reading Direction list

- Left_To_Right
- Right_To_Left
- Top_To_Bottom
- Bottom_To_Top

8. Text Type List:

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

9. Colour approximation list:

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

10. 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)