Image and Text Extractor User Guide

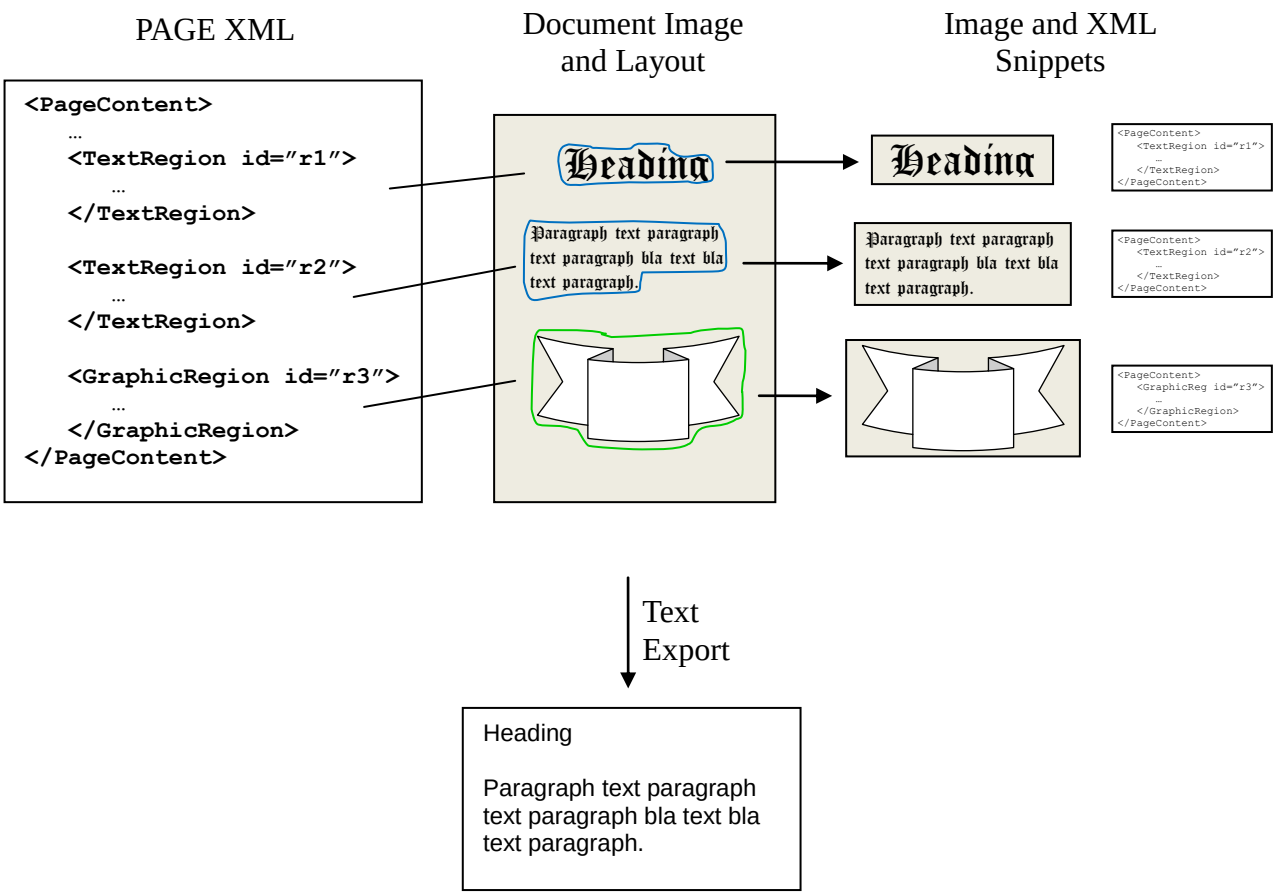
Version 1.1
Date: November 2011

Contents

1 About Image and Text Extractor

The command line tool can be used to extract document snippets (image / layout description) for layout elements of document layouts in PAGE XML format (see publications at http://www.primaresearch.org). Furthermore, the text content of layout regions can be serialised according to the reading order and exported into a text file.

Example:



2 Using the Image and Text Extractor

Command Line Syntax:

Extractor.exe <filter> <filter value> <image> <xml> <output folder> [options] [params]

Where:

<filter> one of:

"type" - to extract all layout elements of the specified type
"id" - to extract the layout element with the specified ID

<filter value>

For filter "type":

"text" - Text regions. Add a comma separated list of sub-types in brackets to filter by sub-type. E.g.:
"text(paragraph,heading,footnote)"
"textline" - Text lines
"word" - Words
"glyph" - Glyphs
"image" - Image regions
"linedrawing" - Line drawing regions
"graphic" - Graphic regions
"table" - Table regions
"chart" - Chart regions
"separator" - Separator regions
"maths" - Maths regions
"noise" - Noise regions
"frame" - Frame regions
"unknown" - Unknown regions
"border" - Document border

For filter "id":

The ID of the element to be extracted (region, text line, word or glyph).

<image> The document image file (TIFF format).

<xml> The document layout file (PAGE format).

<output folder> Folder where to store the output files. (Optional: Use "-" for current folder.)

[options] Extraction options. A combination of (no spaces):

x - To generate PAGE XML snippets in addition to image snippets.
b - To use bounding boxes for extracting elements. Usually the polygonal outline of an element is used to extract an image snippet. The background is filled with white. Use this option to cut rectangular instead so that no background has to be filled.

t - To export the text content (only on region level).
Note: This option disables the image extraction.
A placeholder (e.g. '-') may be used for the image file.

[params] Optional parameter file (.ini) for text export (relative paths have to start with .\).
Example:

```
[TextExporter]
UseOnlyRegionsInReadingOrder=1
InsertExtraLineBreakAfterRegions=1
```

Text Export

The text export (option “t”) serialises the text content of all selected regions (specified by the filter value) according to reading order and y-position and saves it to a text file (same name as the XML input file). At the moment it is not possible to export the text of text line, word or glyph elements.

Parameters:

UseOnlyRegionsInReadingOrder (“1” or “0”, default is “0”)

If set to “1” only regions that are part of the logical reading order description are used for exporting the text. Not all text regions are necessary part of the reading order. Page numbers for instance are usually excluded. This option then also excludes these regions from the text output.
When set to “0” all regions are used for the export. However, regions not belonging to the reading order will be appended at the end and ordered according to their vertical position within the document.

InsertExtraLineBreaksAfterRegions (“1” or “0”, default is “1”)

If set to “1” an extra line break is inserted after each text region (the text contents of regions are then separated by empty lines).