Extractor and Exporter User Guide

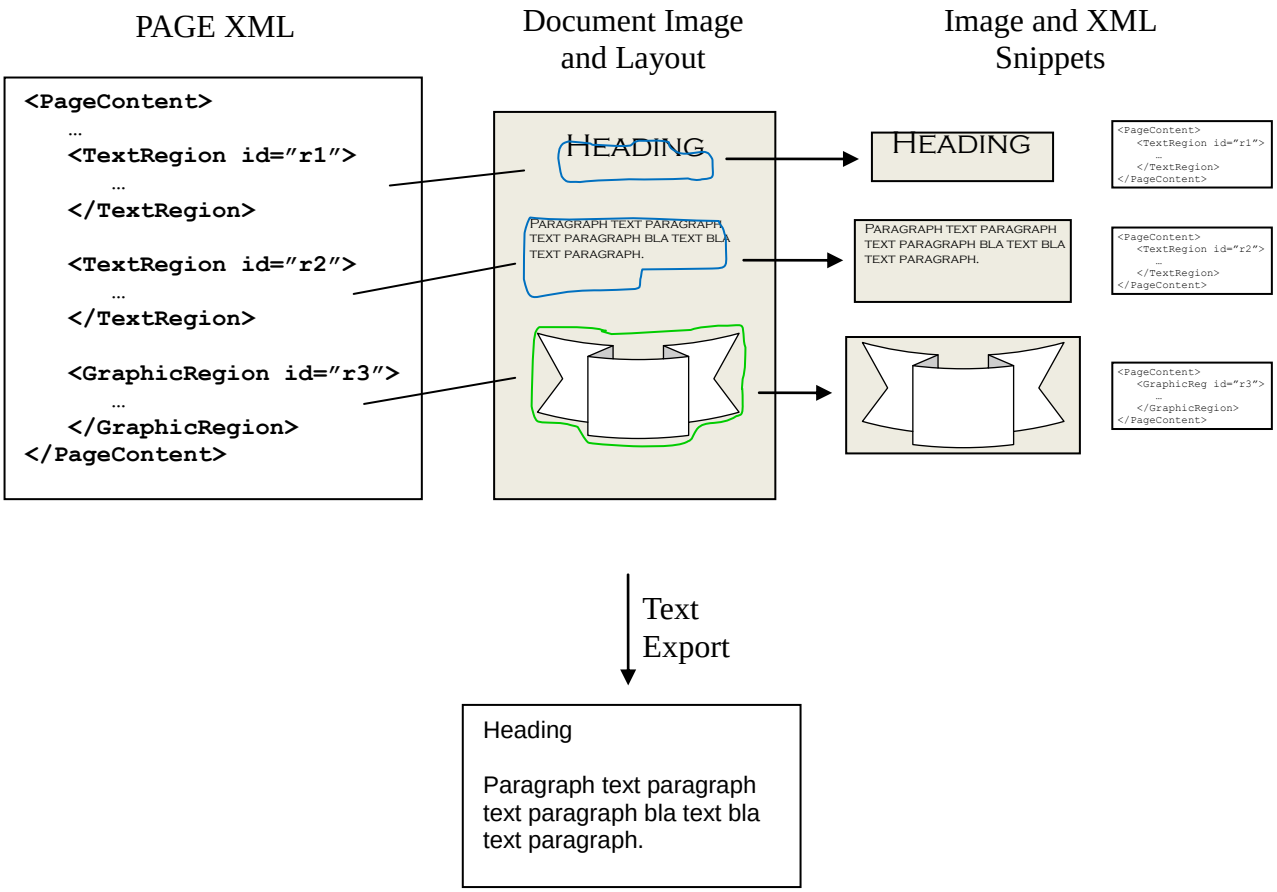
Version 1.2
Date: August 2013

Contents

1 About the Extractor/Exporter

The command line tool can be used to extract document snippets (image / layout description) for layout elements of documents 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 Extractor/Exporter

Command Line Syntax

```
ExtractorExporter <arg1> <val1> <arg2> <val2> ... [opt1] [otp2]

Arguments and options:

  Extracting image and/or layout snippets:

    -extract snippetType1[,snippetType2]
        Supported snippet types:
           imageSnippets
           layoutSnippets
    -filter-by ...
          type - extracts all regions of the specified type
          id   - extracts the region with the specified ID
    -filter ...
       For 'filter-by type':
          text        - text region. Add comma separated a list of sub-types
                    in brackets to filter by sub-type. E.g.:
                      text(paragraph,heading,footnote)
            textline  - text line sub-region
            word      - word sub-region
            glyph     - glyph sub-region
          image       - image region
          linedrawing - line drawing region
          graphic     - graphic region
          table       - table region
          chart       - chart region
          separator   - separator region
          maths       - maths region
          chem        - chemical formula region
          music       - musical notation region
          advert      - maths region
          noise       - noise region
          unknown     - unknown region
          border      - document border
       For 'filter-by id':
          The ID of the page object (region) to be extracted.
    -image <file path> The document image file
    -page-content <file path> The document layout XML file (PAGE).
    -output-folder <folder> Folder where to store the output images.
                            Use '-'for current folder.
    -boxes   To use bounding boxes instead of polygons for extracting
             image snippets

  Exporting text:
    -export text
    -filter region|textline|word|glyph(...) (optional)
          Add comma separated a list of region sub-types
          in brackets to filter by sub-type. E.g.:
            region(paragraph,heading,footnote)
    -page-content <file path> The page content XML file (PAGE).
    -output-folder <folder> Folder where to store the text file.
    -param-file <file path> Parameter file with export settings (optional).
                            (Relative path has to start with .\ ).
                          Example for ini file:
                              [TextExporter]
                              UseOnlyRegionsInReadingOrder=1
                              InsertExtraLineBreakAfterRegions=1
    -ignore-errors   Ignore page content errors.

  Exporting Gamera XML:
    -export gamera
    -page-content <file path> The page content XML file (PAGE) containing glyphs.
    -image <file path> Bi-level document image file
    -output-file <file path> Target file path for the Gamera XML file.
    -param-file <file path> Parameter file with export settings (optional).
                            (Note: Relative paths have to start with .\ ).
    -ignore-errors   Ignore page content errors.
```

Text Export

The text export 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).

Gamera XML Export

The Gamera export creates an XML document with characters for training with the Gamera OCR engine (http://gamera.informatik.hsnr.de/addons/ocr4gamera/).

Following export parameters are supported:

CharacterNameLookUpTableFile (file path, optional)

XML file with look-up table for character naming. By default the file <ExtractorExporter\data\Gamera\ defaultCharacterNames.xml is used.

GroupCharacterClassName (string, optional)

Specifies the string that is prepended for groups broken characters (e.g. latin small letter i). By default nothing is prepended. If for example “_group” is specified the result for letter i would be “_group.latin.small.letter.i” instead of “latin.small.letter.i”.

Example.ini content:

```
[GameraExporter]
CharacterNameLookUpTableFile=c:\test\GameraCharacterLookup.xml
GroupCharacterClassName=_group
```