JPageConverter User Guide

Version 1.5
November 2019

www.primaresearch.org

Overview

JPageConverter is a Java based tool to convert document layout XML files in PAGE format to the latest format version. It also provides basic features to supplement data in PAGE files.

For more information on the PAGE format framework and other software tools see:
www.primaresearch.org/tools

Command Line Interface

```
Arguments:

  -source-xml <XML file>        PAGE XML file to convert.
    OR
  -source-json <JSON file>      Google OCR output file to convert.

  -target-xml <XML file>        Output PAGE XML file.

  -convert-to <schema version>  Target PAGE schema version. (optional)
              Available versions:
                 LATEST
                 2013-07-15
                 2010-03-19

  -set-gtsid <ID|prefix[start,end]>   To set the the GtsId field. (optional)
         Usage:
              Use <ID> to set a specific GtsId.
              Use <prefix[start,end]> to extract the GtsId from the
              filename of -source-xml.
         Examples:
              -set-gtsid 00236178     Given ID
              -set-gtsid pc-[0,7]     Prefix + first 8 characters of filename

  -text-filter <XML file>   Applies filter to the text content. (optional)
  -neg-coords <mode>   Handle negative coordinates (optional)
       Modes:
         removeObj - If an object contains one or more point with negative
                     coordinates, remove the whole object.
         toZero    - Change negative values to 0
```

Examples are provided in the “batch” folder of the JPageConverter installation.

Text Filter

It is possible to apply a text filter to all text elements of the processed document. The text filter contains replacement rules to replace single Unicode characters or a sequence of characters by

another character or sequence. To apply a filter use -text-filter in the command line call and provide the file name of the XML file containing the filter rules. The XML file must have the following format:

```
<?xml version="1.0" encoding="utf-8"?>
<Parameters>
  <Parameter type="4" name="Replacement Rule"
             id="1"
             sortIndex="1"
             value="0065:=0061"
             visible="false"
             isSet="true">
    <Description>Replace e by a</Description>
  </Parameter>
  <Parameter type="4" name="Replacement Rule"
             id="2"
             sortIndex="2"
             value="0074:="
             visible="false"
             isSet="true">
    <Description>Delete all t</Description>
  </Parameter>
</Parameters>
```

Rules designed for the “Page Converter and Validator” tool for Windows can also be used for this tool.
Each parameter element contains a replacement rule. The sortIndex attribute specifies in which order the rules will be applied. The id attribute must be unique (easiest to use the same value as the sort index). The description is optional but helps to understand the rules. The actual rule is encoded in the value attribute. The general format is “HHHH[,HHHH,...]:=[HHHH,HHHH,...]”. HHHH is a Unicode character represented as 4 digit hexadecimal number. In the example above “0065:=0061” means ‘replace all characters e with character a’. To replace a character sequence separate the single characters by comma. The same applies for the right-hand side (the replacement character or sequence). It is also possible to remove characters by leaving the right-hand side empty (e.g. “0074:=” to delete all ts).
For an example how to use a filter, see the example ‘Text normalisation’ in the batch folder of the tool. The example includes an XML file with replacement rules and a batch file to process a document.

Special Rules

There is a set of special rules for more complex replacements that cannot be described by the default rule syntax. These rules are applicable by using predefined keywords on the left side of a rule (before the :=). See following table for the keywords and their function:

Keyword for Special Rule	Description

MULTSPACE	Stands for a variable number of consecutive spaces (two or more)
STARTSPACE	Stands for a space at the beginning of a text
ENDSPACE	Stands for a space at the end of a text
MULTBREAK	Stands for a variable number of consecutive line breaks (two or more)
STARTBREAK	Stands for a line break at the beginning of a text
ENDBREAK	Stands for a line break at the end of a text
REGEX####	Generic regular expression (see Java documentation). Note: This rule is not supported by the "Page Converter and Validator" tool for Windows.

Examples:
"_MULTSPACE_:=0020" Replaces all multiple spaces by one space
"_STARTSPACE_:=" Removes a space at the beginning of a text
"_REGEX_\p{Punct}:=" Removes all punctuations

Page Element Filter

Rules can be restricted to individual text element levels (text region, text line, word and glyph) by adding a filter rule at the end of a replacement rule. The extended rule format is defined as: "HHHH[,HHHH,...]:=[HHHH,HHHH,...][|[R][L][W][G]]". Thus, a text element filter can be applied by adding "|" followed by a combination of "R" (region), "L" (text line), "W" (word) and "G" (glyph). The text replacement rule will be carried out only for the specified text element levels.

Example: "0020:=|WG" Removes all spaces from the text of words and glyphs