2013 PAGE XML Format Update

What you need to know

Contents

19 November 2013
PRImA Research Lab
University of Salford
United Kingdom
www.primaresearch.org

What is PAGE?

The PAGE (Page Analysis and Ground truth Elements) format framework incorporates several XML schemas representing the whole workflow of document analysis, including image enhancement, binarisation, geometrical correction, layout analysis, layout evaluation and OCR. This document focuses on the XML schema for document layouts, which allows for polygonal regions with various attributes (including text content), reading order, layers and more.

For more information see this publication: www.primaresearch.org/publications/PAGE

What has changed?

More compact

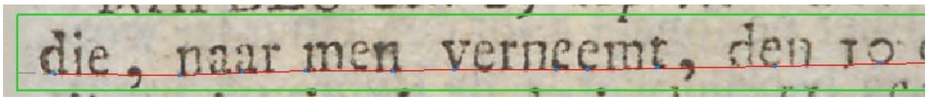
In accordance with other major document layout formats we made the representation of polygons more compact. This leads to considerably smaller file sizes and shorter loading as well as writing times.

More robust

The 2013 schema is stricter which, in combination with validation, makes PAGE XML and tools that use it more robust.

More features

Following table describes the most important new features. For a complete list see the change log.

Nested regions	All regions can now have nested sub-regions. Before this was limited to *FrameRegion* elements. For instance, a table region can now have text and/or image regions for the table cells. Note that the *FrameRegion* has been replaced by *GraphicRegion* with subtype “frame”.
New region types	*MusicRegion* for musical notations, *ChemRegion* for chemical formulas and *AdvertRegion* for advertisements.
New attributes	Text style (font family, bold, italic, underlined, ...) Text production (printed, handwritten, typewritten, ...) Page type (front cover, title, index, table of contents, blank, ...)
Baselines	Baselines can now be defined for text lines. 
New relations	New *Relation* element to model relations between layout objects (e.g. drop-capital - paragraph, image - caption).
Custom fields	Most layout elements now have attributes for generic content (“comments” and “custom”).
...	

What do I do with my document layout files in 2010 PAGE format?

Not to worry, all tools developed at the PRImA research lab support all previous versions of the PAGE format. Nevertheless, if you would like to profit from the more compact 2013 format, you can convert all your files in one go using *Page Converter and Validator for Windows* or *JPageConverter* (Java). Both tools come with script files for batch conversion.

Additional Information and Resources

2013 Schema for PAGE XML format:

http://schema.primaresearch.org/PAGE/gts/pagecontent/2013-07-15/pagecontent.xsd

Tools supporting PAGE by PRImA:

http://www.primaresearch.org/tools

The PRImA research group itself:

http://www.primaresearch.org

University of
Salford
MANCHESTER

PRImA
Research Lab