Proceedings of the 20th International Conference on Pattern Recognition (ICPR2010), Istanbul, Turkey, August 23-26, 2010, IEEE-CS Press, pp. 257-260.

The PAGE (Page Analysis and Ground-truth Elements) Format Framework[†]

S. Pletschacher and A. Antonacopoulos
Pattern Recognition and Image Analysis (PRImA) Research Lab
School of Computing, Science and Engineering, University of Salford, United Kingdom
http://www.primaresearch.org

Abstract

There is a plethora of established and proposed document representation formats but none that can adequately support individual stages within an entire sequence of document image analysis methods (from document image enhancement to layout analysis to OCR) and their evaluation. This paper describes PAGE, a new XML-based page image representation framework that records information on image characteristics (image borders, geometric distortions and corresponding corrections, binarisation etc.) in addition to layout structure and page content. The suitability of the framework to the evaluation of entire workflows as well as individual stages has been extensively validated by using it in high-profile applications such as in public contemporary and historical ground-truthed datasets and in the ICDAR Page Segmentation competition series.

1. Introduction

There is a significant need to explicitly record information of diverse nature used or produced by methods within digitization pipelines – both for making the substitution of alternative methods possible and for evaluating intermediate stages.

A format framework that would fulfil this necessity should be capable of providing detailed and accurate information on the results of each individual processing step as well as cumulative information on the results of processes that took place so far at any given point in the pipeline.

Furthermore, large scale evaluation requires a common format for both ground truth and results in order to obtain comparable structures which can be processed by automated evaluation tools.

A number of existing formats (e.g. ALTO [1] or hOCR [2]) have been developed predominantly for recording analysis and recognition results. Such formats are well suited to store final outcomes of a document analysis pipeline (e.g. text and associate attributes) and to feed publication systems. However, due to specific concepts on how recognised elements are represented, such formats cannot always be used in the detailed evaluation of individual processing steps (e.g. ALTO has no concept of words or glyphs but only strings, which renders it impractical for the evaluation of glyph and word segmentation).

On the other hand there are formats solely devoted to specific evaluation approaches for one processing step or type of methods (e.g. segmentation evaluation based on labelled pixel maps, bounding boxes etc. [3], [4]). While such formats are very useful for detailed evaluation they have to be created in addition to the actual desired output and cannot be used for representing the structure or the content of analysed documents.

This paper presents a format framework which has been specifically designed to integrate useful features of existing formats but also to overcome some of their fundamental limitations. Its flexible structure and ability to express dependencies between processing steps predestines it for performance evaluation not only of individual methods but also whole workflows.

2. Design

A major goal of the proposed PAGE (Page Analysis and Ground-truth Elements) format framework is to harmonise the requirements for storage of intermediate processing results as well as performance evaluation

[†] This work has been supported in part through the EU 7th Framework Programme grant IMPACT (Ref: 215064).

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.72

aspects to aid the assessment of document image analysis methods on **page level**. To achieve this, it is necessary to represent results as well as ground truth in a way that allows direct comparison of corresponding elements, ideally by making use of the same format. Therefore, such a format must support a highly detailed and accurate description of any information which can be derived from a given document image while still being valid in case an analysis method does not support the whole range of features.

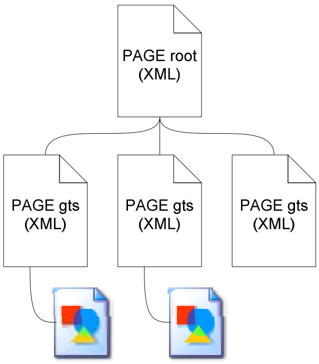


Figure 1. PAGE structure consisting of a root instance linking to task-specific output or ground truth files. Note that gts XML instances may link to further resources (e.g. a dewarped image) depending on the nature of the respective method.

Another important aspect is the **partitioning** of data. Especially for the transition from evaluating individual methods to whole process chains, it is necessary to have a flexible structure which supports and keeps track of different workflow configurations and respective interim results. This is a feature which has not been considered in previous formats. Instead of having one integrated result file, the PAGE format framework introduces a two-tier architecture: A root structure manages results of multiple processing steps by resolving **dependencies** and linking to the actual task specific data instances. In the same way it is possible to describe dependencies between ground truth instances. The decoupling of different ground truth subsets makes it possible to express not only dependencies but also different views. A scenario for this is the evaluation of layout analysis in both the original and a dewarped version of this image. As coordinates of regions may have changed after dewarping, a second ground-truth subset (i.e. a different view for still the same evaluation task) which depends on perfect dewarping of the original (i.e. the dewarping ground truth subset) can be linked. Another example for different views is flat text vs. structured layout including text in text line, word and glyph elements.

3. Format components

As outlined before, the PAGE format framework consists of two tiers: A **root** instance to reflect the structure of a particular workflow or evaluation experiment and **gts** (ground truth and storage) instances containing the actual data for the processing steps involved (see Figure 1). The format of the root structure as well as all task specific sub-formats are specified by means of separate XML Schema definitions which are publicly available [5]. Currently, the specific formats cover preprocessing steps like binarisation, deskew and dewarping as well as a sophisticated page content description relating to document analysis and recognition methods.

3.1 Linking root structure

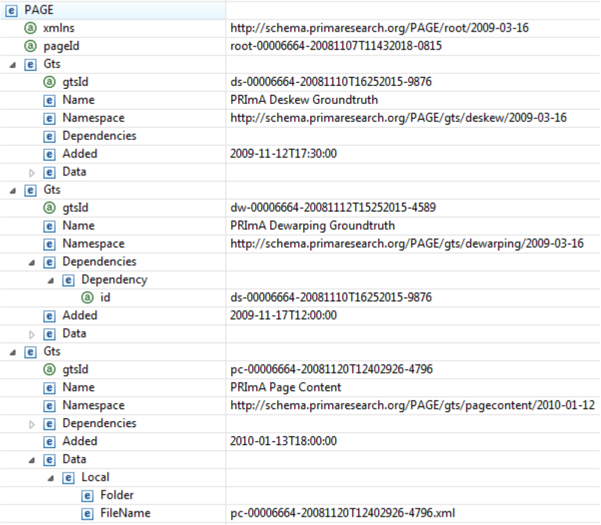
The need for a root structure results from the design decision to separate the data of distinct methods in order to have both access to interim results and more flexibility in terms of workflow configurations including alternative process chains. Accordingly, the root structure contains links to all individual processing results (in evaluation scenarios these would be ground truth sets) including information about the employed method and output format (for an example see Figure 2).

Formats are uniquely identified via their namespaces and every *gts* entry can be referenced by its *id*. This way, dependencies between instances can be expressed using the standard XML mechanism of ID and IDREF. Dependencies arise typically if the input required for one method is the output of another. Image segmentation, for instance, can be carried out on a binarised image or on a first dewarped and then binarised image. The absolute coordinates of detected regions in the two cases will be obviously different which illustrates the importance of recording such dependencies. Likewise, to evaluate the two results from the above example, segmentation ground truth has to be created for the original image as well as the dewarped image.

In terms of the format, there is no difference in representing *results* in production scenarios and *ground truth* in evaluation scenarios. The only difference is that ground truth is typically created manually and considered the correct (ideal) output for a particular task.

3.2 Image preprocessing data formats

The PAGE framework currently comprises formats for binarisation, deskewing and dewarping in order to make interim results (or ground truth) for these document analysis related preprocessing steps accessible. According to their nature, the output of such methods is, like the input, an image. Hence, all preprocessing formats allow links to further external files representing the actual output (or ground truth) in addition to parameters and transformation model data where applicable. They all contain, like any *gts* sub-format, a *metadata* section and an *id*.



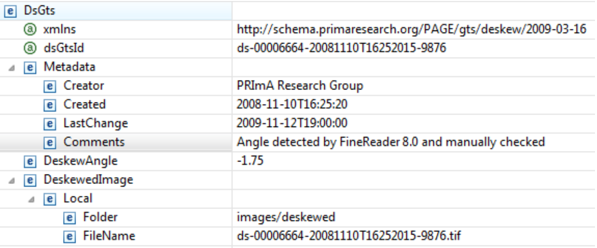
PAGE	
xmlns	http://schema.primaresearch.org/PAGE/root/2009-03-16
pageId	root-00006664-20081107T11432018-0815
Gts	
gtsId	ds-00006664-20081110T16252015-9876
Name	PRImA Deskew Groundtruth
Namespace	http://schema.primaresearch.org/PAGE/gts/deskew/2009-03-16
Dependencies	
Added	2009-11-12T17:30:00
Data	
Gts	
gtsId	dw-00006664-20081112T15252015-4589
Name	PRImA Dewarping Groundtruth
Namespace	http://schema.primaresearch.org/PAGE/gts/dewarping/2009-03-16
Dependencies	
Dependency	
id	ds-00006664-20081110T16252015-9876
Added	2009-11-17T12:00:00
Data	
Gts	
gtsId	pc-00006664-20081120T12402926-4796
Name	PRImA Page Content
Namespace	http://schema.primaresearch.org/PAGE/gts/pagecontent/2010-01-12
Dependencies	
Added	2010-01-13T18:00:00
Data	
Local	
Folder	
FileName	pc-00006664-20081120T12402926-4796.xml

Figure 2. Schematic view of a PAGE root instance.

Binarisation is supported for simple thresholding as well as skeleton-based methods (e.g. as described in [6]). Therefore, the format allows linking a binarised image plus the corresponding skeleton. The format offers also to revert to image patches within the original as the process of manually creating this kind of binarisation ground truth for whole images is extremely time consuming and costly. To achieve this, regions of interest can be defined (e.g. only text regions in a document) for which the actual data is available.

For global **deskew** methods there is mainly the deskew angle in terms of image transformation (defined as the angle an image has to be rotated in clockwise direction in order to correct the present skew; negative values indicate anti-clockwise rotation) and the deskewed image as output (see Figure 3).

Dewarping model data can potentially be very complex if apart from page curl also arbitrary warping effects are to be considered. In the current version, the format facilitates besides linking of dewarped images also grid models for more sophisticated evaluation approaches (e.g. like [7]).



DsGts	
xmlns	http://schema.primaresearch.org/PAGE/gts/deskew/2009-03-16
dsGtsId	ds-00006664-20081110T16252015-9876
Metadata	
Creator	PRImA Research Group
Created	2008-11-10T16:25:20
LastChange	2009-11-12T19:00:00
Comments	Angle detected by FineReader 8.0 and manually checked
DeskewAngle	-1.75
DeskewedImage	
Local	
Folder	images/deskewed
FileName	ds-00006664-20081110T16252015-9876.tif

Figure 3. Schematic view of a PAGE deskew instance.

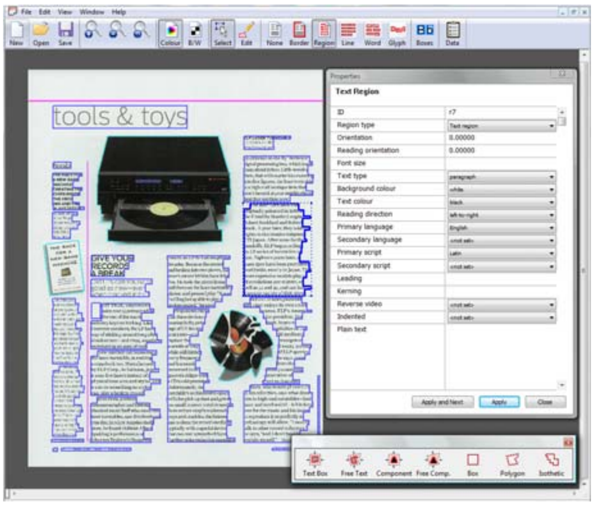
3.3 Page content data format

The page content gts format is currently the most deep sub-format of the PAGE framework. It allows precise description of any content elements which can be found in document images. On the highest level it is possible to specify the document border in order to mark background not belonging to the page which might be present in an image due to the scanning process.

Within the page, all elements are considered to be in a region of a specific *type*. The most important region types are *text*, *image*, *line drawing*, *graphic*, *table*, *chart*, *separator*, *maths*, *noise* and *frame*. Moreover, it is possible to mark regions as *unknown*. Text regions can be further specified by means of *text lines*, *words* and *glyphs*.

For each element representing a region on the page there is a description of its outline in the form of a polygon defined by its *coordinates*. Besides the outline and an *id* which is common for all regions there is also a set of region type specific metadata. Text regions, for instance, may contain information about *language*, *script*, *font*, *reading direction*, *text colour*, *background colour*, *type* (e.g. heading, paragraph, caption, footer, etc.) among others.

As text recognition is one of the major concerns in document analysis, it is possible to store ASCII or Unicode encoded text for all textual elements. Enabling text on all levels (from glyphs, words, text lines up to text regions) carries the risk of redundant data. Nevertheless, this is necessary in order to achieve a maximum of flexibility. While text ground truth, for instance, might be available on glyph level, a specific OCR system might output text only for whole text regions. Evaluation in such cases is only possible on the highest common level (text from lower levels can be pulled up).



Text Region

ID	r7
Region type	Text region
Orientation	0.00000
Reading orientation	0.00000
Font size	
Text type	paragraph
Background colour	white
Text colour	black
Reading direction	left-to-right
Primary language	English
Secondary language	<not set>
Primary script	Latin
Secondary script	<not set>
Leading	
Kerning	
Reverse video	<not set>
Indented	<not set>
Plain text	

Figure 4. Graphical example of a page content instance displayed in the current version of Aletheia, a ground-truthing tool supporting PAGE.

As for reading order, the format offers more flexibility than other existing formats. Apart from defining strictly linear relations it is also possible to assign groups of ordered or unordered elements which may also be nested. This is particularly useful for complex layouts as can be found in newspapers and magazines. Another sophisticated feature is the layer concept which can be used to specify the order of superimposed and overlapping regions. Both, reading order and layers, are stored only as logical constructs referencing the particular regions by means of their *ids*.

The format for page content is used by the PRImA layout analysis evaluation system [8].

4. Concluding remarks

The PAGE format framework described in this paper has been developed over the last years and has reached a mature state. It is used in high-profile applications such as evaluation datasets for layout analysis of contemporary documents [9], datasets and extensive evaluation infrastructure for historical documents (within the scope of the IMPACT project) as well as international competitions (ICDAR competition series) [10].

The use of XML as basic technology allows straightforward integration into most infrastructures – not only for research projects but also in production environments. Like ALTO, it can be used as an extension schema to METS (Metadata Encoding and Transmission Standard) which is commonly used to store metadata and structural information in library environments. Apart from its use in tailor-made solutions there is also a complete performance evaluation infrastructure supporting it (tools for semi-automatic ground-truthing, searchable online datasets, evaluation metrics and scenarios, performance evaluation tools, viewers, converters, I/O libraries etc.).

Finally, the format is also providing for future extensions. New *gts* sub-formats can be devised if an application or evaluation scenario requires it (e.g. for font recognition). The root structure can be used to point to any kind of data instance which is identified by a namespace and defined by a corresponding XML Schema. Versions of already existing sub-formats (if future developments require amendments or changes) are distinguished through unique namespace identifiers which include the publication date (e.g. http://schema.primaresearch.org/PAGE/gts/pagecontent/2010-03-19).

References

[1] ALTO - Analyzed Layout and Text Object, as of version 2.0 maintained by the Library of Congress, http://www.loc.gov/standards/alto/

[2] T. Breuel, “The hOCR Microformat for OCR Workflow and Results”, *Proc. ICDAR2007*, pp. 1063-1067.

[3] B. Yanikoglu and L. Vincent, “Pink Panther: A Complete Environment for Ground-Truthing and Benchmarking Document Page Segmentation. *Pattern Recognition*, vol. 31, pp. 1191–1204, 1998.

[4] B. Gatos, A. Antonacopoulos and N. Stamatopoulos, “ICDAR2007 Handwriting Segmentation Contest”, *Proc. ICDAR2007, pp. 1284-1288.*

[5] PAGE - Page Analysis and Ground-truth Elements, http://schema.primaresearch.org/PAGE/

[6] K. Ntirogiannis, B. Gatos, I. Pratikakis, “An Objective Evaluation Methodology for Document Image Binarization Techniques”, *Proc. DAS2008*, pp. 217-224.

[7] N. Stamatopoulos, B. Gatos, I. Pratikakis, “A Methodology for Document Image Dewarping Techniques Performance Evaluation”, *Proc. ICDAR2009*, pp. 956-960.

[8] A. Antonacopoulos and D. Bridson, “Performance Analysis Framework for Layout Analysis Methods”, *Proc. ICDAR2007*, pp. 1258-1262.

[9] A. Antonacopoulos, D. Bridson, C. Papadopoulos and S. Pletschacher, “A Realistic Dataset for Performance Evaluation of Document Layout Analysis”, *Proc. ICDAR2009*, pp. 296-300.

[10] A. Antonacopoulos, S. Pletschacher, D. Bridson, C. Papadopoulos, “ICDAR2009 Page Segmentation Competition”, *Proc. ICDAR2009*, pp. 1370-1374.