# Layout Evaluation User Guide

*Version 1.8*
*Date: July 2017*

## Contents

# System Requirements

### Minimum System Requirements

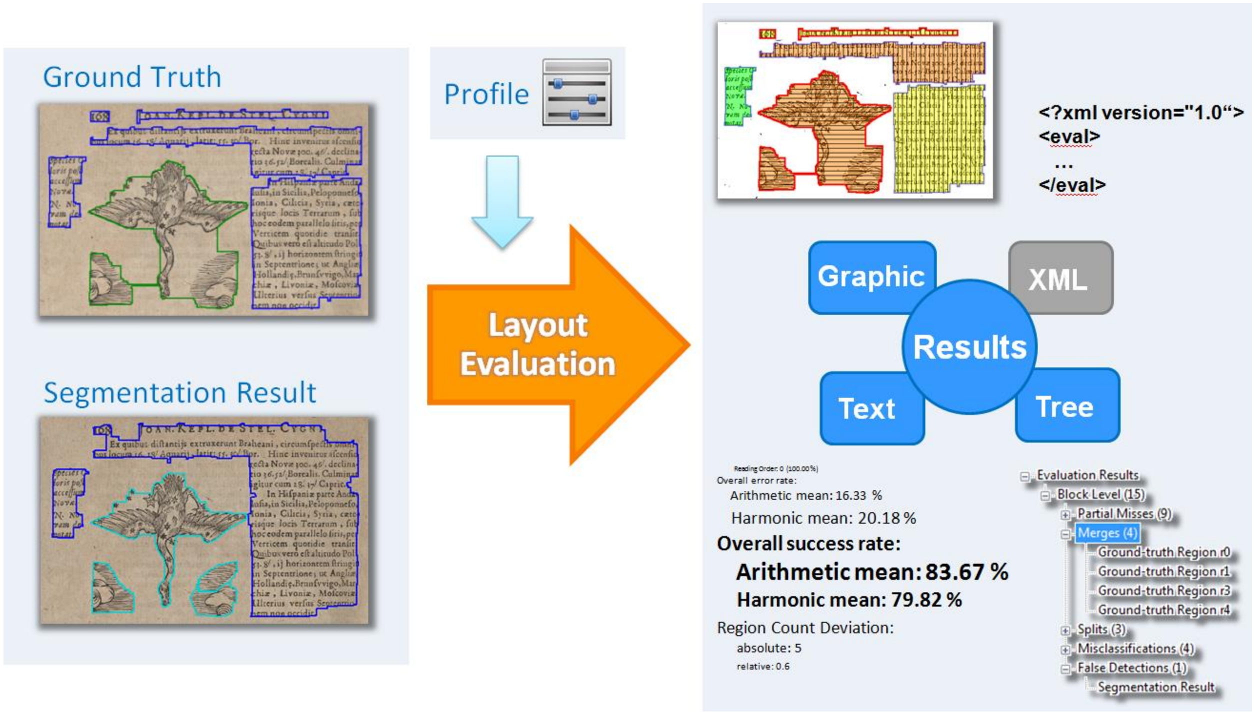
| | |
|---|---|
| Operating System: | Windows 7 or higher |
| CPU: | 1.5 GHz |
| RAM: | 2 GB |
| Screen Resolution: | 1024 * 768 |
| Hard disk space: | 500 MB |

### Recommended System Configuration

| | |
|---|---|
| Operating System: | Windows 7 or higher |
| CPU: | 3.0 GHz dual Core |
| RAM: | 4 GB |
| Screen Resolution: | 1600 * 1200 |
| Hard disk space: | 250 GB |

# Introduction

Layout evaluation is used to benchmark results of layout segmentation methods and uncover specific problems of the algorithms to help developers to improve them. As input the ground truth XML file, the segmentation result XML file and the black-and-white document image are required. The colour image is optional and only used for viewing. For the evaluation the ground truth regions are compared to the segmentation result regions. Differences are logged as evaluation errors. Weights and settings for a specific scenario can be specified using an evaluation profile. See the following illustration as a general overview:



```
<?xml version="1.0“>
<eval>
  …
</eval>
```

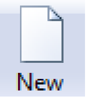
Reading Order: 0 (100.00%)
Overall error rate:
Arithmetic mean: 16.33 %
Harmonic mean: 20.18 %
**Overall success rate:**
**Arithmetic mean: 83.67 %**
**Harmonic mean: 79.82 %**
Region Count Deviation:
absolute: 5
relative: 0.6

```
⊟ Evaluation Results
  ⊟ Block Level (15)
    ⊞ Partial Misses (9)
    ⊟ Merges (4)
        Ground-truth Region r0
        Ground-truth Region r1
        Ground-truth Region r3
        Ground-truth Region r4
    ⊞ Splits (3)
    ⊞ Misclassifications (4)
    ⊟ False Detections (1)
        Segmentation Result
```

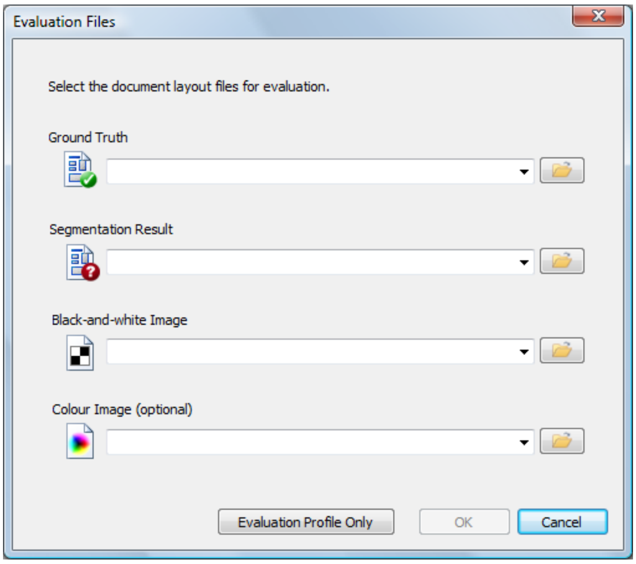
# Using the Tool

## Preparing a new Evaluation

Click the ‘New’ tool bar button.



A dialogue pops up:



Evaluation Files

Select the document layout files for evaluation.

Ground Truth

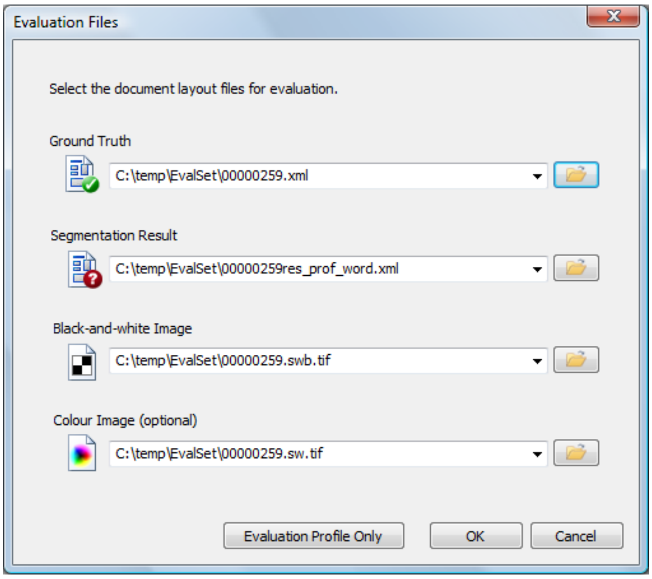
Segmentation Result

Black-and-white Image

Colour Image (optional)

Click the buttons on the right to select the ground truth XML file, the segmentation result XML file and the image files. If a ground truth file has been selected, the drop-down box for the segmentation result will be populated with files having the same document ID (a number). If the right document is amongst the suggestions, it can be selected straight away, without using the ‘File Open’ dialogue.

Once all required files are chosen, the OK button will be enabled:



Evaluation Files

Select the document layout files for evaluation.

Ground Truth
C:\temp\EvalSet\00000259.xml

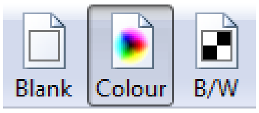
Segmentation Result
C:\temp\EvalSet\00000259res_prof_word.xml

Black-and-white Image
C:\temp\EvalSet\00000259.swb.tif

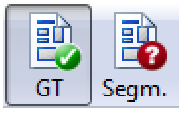
Colour Image (optional)
C:\temp\EvalSet\00000259.sw.tif

# Viewing the Document Layouts

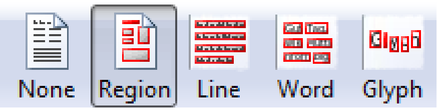
Select which document image to view by clicking the corresponding toolbar button. When ‘Blank’ is selected, a white background will be displayed instead an image.



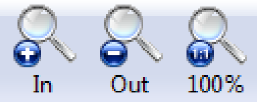
Choose a layout to view by clicking ‘GT’ for ground truth or ‘Segm.’ for segmentation result in the toolbar. It is also possible to select both layouts at once.

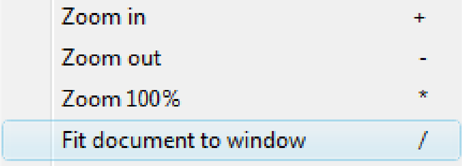


To switch between page element levels use the following toolbar buttons:

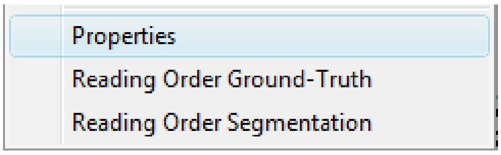


To change the zoom level, use the toolbar buttons, the options in the ‘View’ menu or the mouse wheel while pressing CTRL.

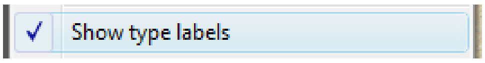


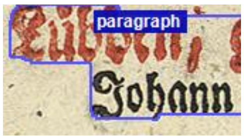


Region properties (attributes) and the reading order can be viewed using the dialogues accessible through the ‘View’ menu:



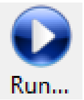
By default the type of a region is displayed as a label in the document view. The labels can be deactivated in the ‘View’ menu:

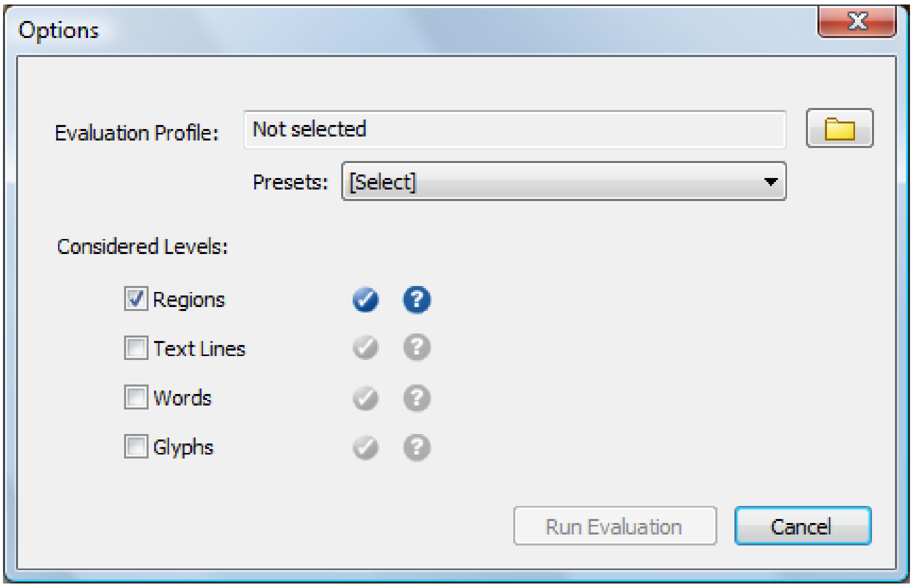




## Running an Evaluation

Open the evaluation dialogue by clicking ‘Run...’ in the toolbar:





Options

Evaluation Profile: Not selected

Presets: [Select]

Considered Levels:

☑ Regions ✓ ?

□ Text Lines

□ Words

□ Glyphs

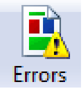
Next, choose an evaluation profile by selecting a preset from the drop-down box or by opening a previously saved profile using the top right button.

Tick the checkboxes for the page element levels that are to be evaluated. By default only the levels are preselected that have data in both the ground truth and the segmentation result. The icons on the right show which data is available. Available ground truth data is displayed as blue tick and available segmentation result data is displayed as blue question mark.

Click ‘Run Evaluation’ to start the evaluation process.

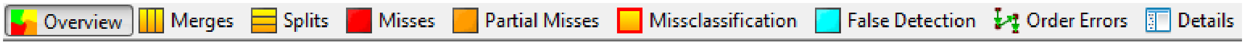
## Browsing Evaluation Results

To view evaluation results, switch to the desired region level and click the ‘Errors’ toolbar button:

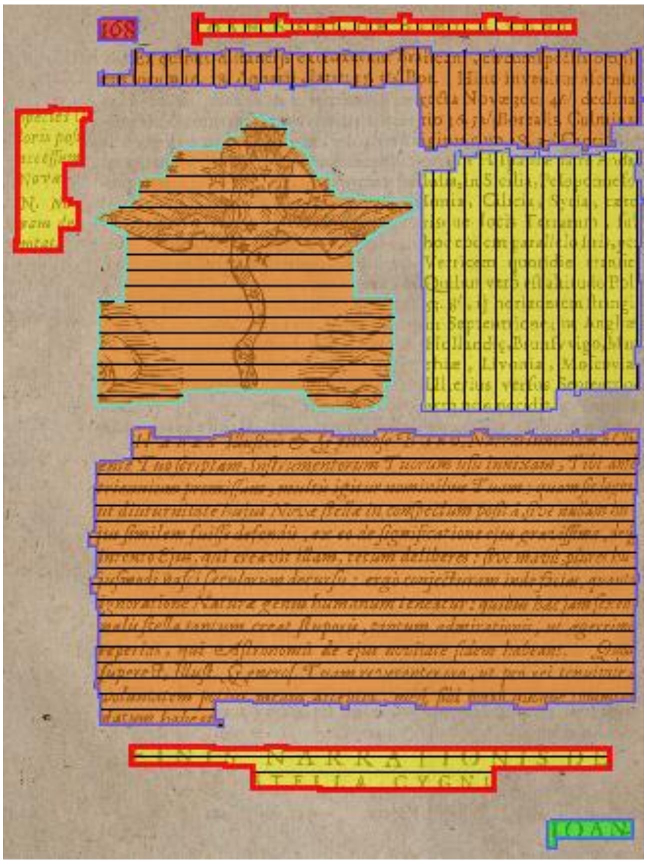


Note: After an evaluation has finished, the tool will automatically switch to the first region level with evaluation results and display the errors.

An additional toolbar is displayed:



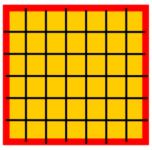
The ‘Overview’ shows all error types at once within the document view:



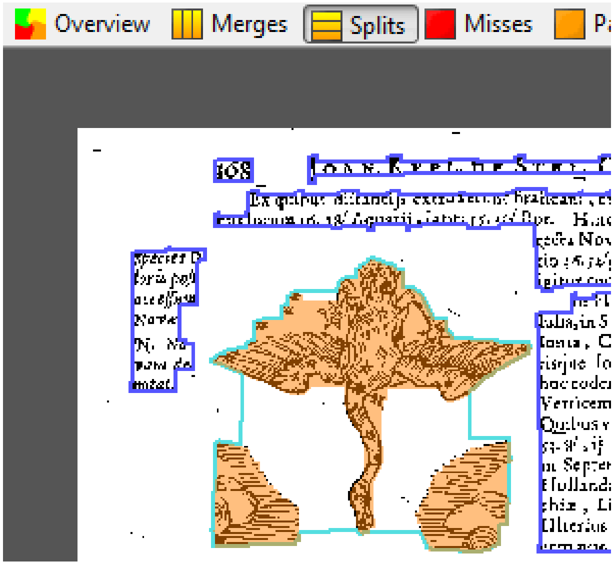
The toolbar is also a legend. The evaluation errors are displayed as follows:

- ▥ Merge
- ▤ Split
- 🟥 Miss
- 🟧 Partial Miss
- 🟨 Misclass.
- 🟦 False Detect.
- 🟩 Correct

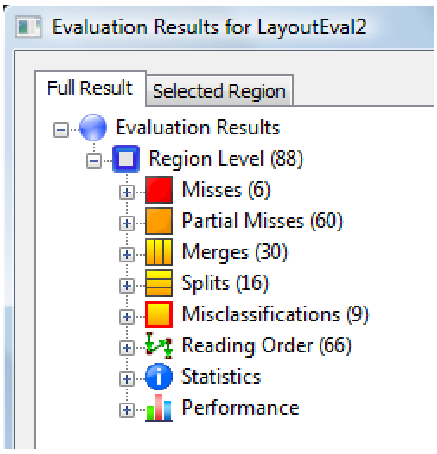
For one region there can also be different errors at once. The colours and patterns are designed that way. If for instance a region is merged, split, misclassified and incomplete, then the error would be highlighted as:



To view the errors by type, switch between the modes ‘Merges’, ‘Splits’, ‘Misses’, ‘Partial Misses’, ‘Misclassification’ and ‘False Detection’. In these modes, the errors are always highlighted as orange regions.

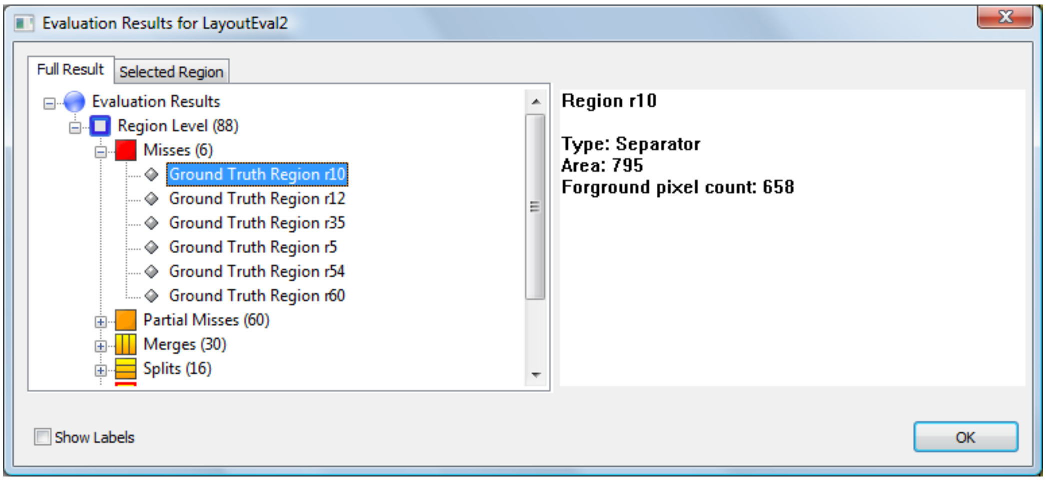


For detailed textual results and a performance overview click ‘Details’ in the toolbar. A dialogue opens:

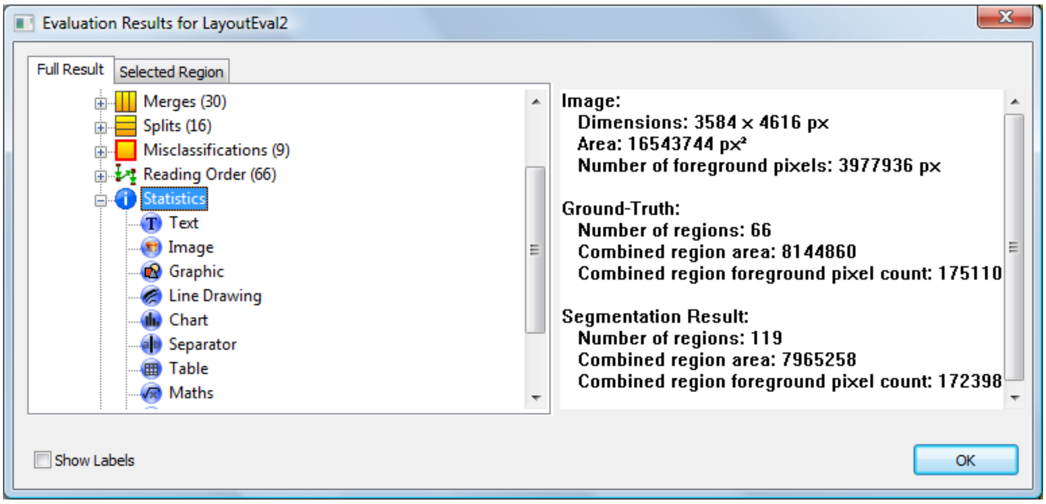


- Evaluation Results
  - Region Level (88)
    - Misses (6)
    - Partial Misses (60)
    - Merges (30)
    - Splits (16)
    - Misclassifications (9)
    - Reading Order (66)
    - Statistics
    - Performance

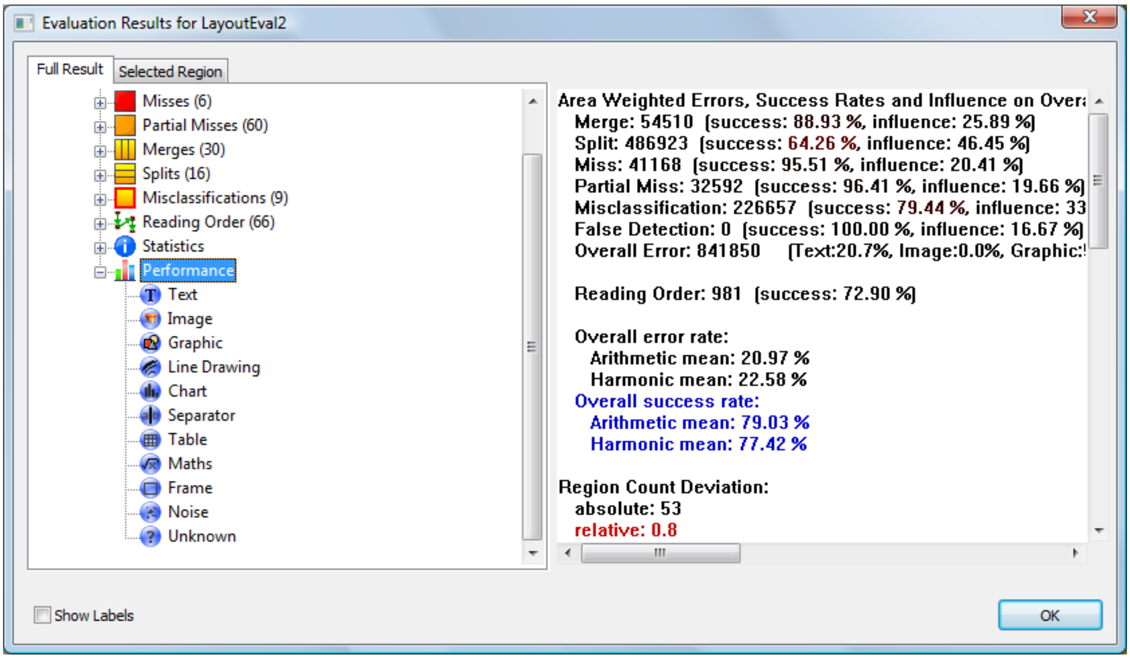
The dialogue contains a tree for all evaluation results. The tree is arranged by page element level (region, text line, word and glyph). For each level there are items for evaluation errors (measures), statistics and performance (success rates, ...). Click on a node to see additional information in the pane on the right. Expand the error nodes to see the involved regions. Expand the statistics or metrics nodes to see results per region type (text, image, table, ...).



```
Region r10

Type: Separator
Area: 795
Forground pixel count: 658
```

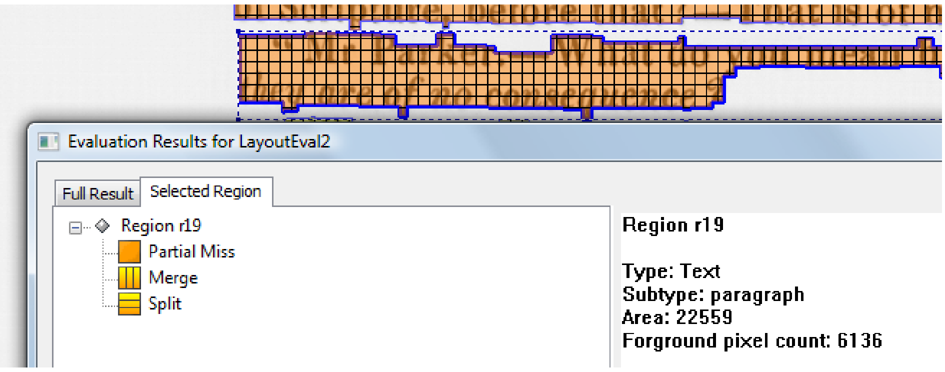


```
Image:
  Dimensions: 3584 x 4616 px
  Area: 16543744 px²
  Number of foreground pixels: 3977936 px

Ground-Truth:
  Number of regions: 66
  Combined region area: 8144860
  Combined region foreground pixel count: 175110

Segmentation Result:
  Number of regions: 119
  Combined region area: 7965258
  Combined region foreground pixel count: 172398
```



```
Area Weighted Errors, Success Rates and Influence on Overa
  Merge: 54510  (success: 88.93 %, influence: 25.89 %)
  Split: 486923  (success: 64.26 %, influence: 46.45 %)
  Miss: 41168  (success: 95.51 %, influence: 20.41 %)
  Partial Miss: 32592  (success: 96.41 %, influence: 19.66 %)
  Misclassification: 226657  (success: 79.44 %, influence: 33
  False Detection: 0  (success: 100.00 %, influence: 16.67 %)
  Overall Error: 841850     (Text:20.7%, Image:0.0%, Graphic:!

  Reading Order: 981  (success: 72.90 %)

  Overall error rate:
    Arithmetic mean: 20.97 %
    Harmonic mean: 22.58 %
  Overall success rate:
    Arithmetic mean: 79.03 %
    Harmonic mean: 77.42 %

Region Count Deviation:
  absolute: 53
  relative: 0.8
```

To view details for a specific region, switch to the ‘Selected Regions’ tab and select a region in the main document view or simply double click the region of interest. A tree shows all segmentation errors found for the selected region. Click the tree items to display details.



Region r19

Type: Text
Subtype: paragraph
Area: 22559
Forground pixel count: 6136

### Evaluation Statistics and Performance Indexes

Based on the raw data, the actual error values and success rates are calculated. The metrics are calculated for a specified structure level (region, text line, word or glyph).
Following figures are produced:

- Statistics
  - Image Area
  - Number of black pixels within the image
  - Overall number of regions (ground-truth and segmentation result)
  - Number of regions per type (text, image, ...) (ground-truth and segmentation result)
  - Overall region area (ground-truth and segmentation result)
  - Overall number of black pixels in regions (ground-truth and segmentation result)
  - Glyph / character OCR error details
- Performance Indexes
  - Overall weighted area error per error type (merge, split, ...)
  - Overall weighted area error per region type (text, image, ...)
  - Overall weighted area error
  - Weighted area success rate per error type
  - Overall weighted count error per error type (merge, split, ...)
  - Overall weighted count error per region type (text, image, ...)
  - Overall weighted count error
  - Weighted count success rate per error type
  - Reading order error
  - Reading order success rate
  - Overall weighted area success rate (arithmetic mean, harmonic mean)
  - Overall weighted count success rate (arithmetic mean, harmonic mean)
  - Recall per type
  - Recall (strict / non-strict)
  - Precision per type
  - Precision (strict / non-strict)
  - F-Measure (strict / non-strict)
  - Region count deviation (absolute / relative)

The same figures are also calculated for each region type (text, image, ...) separately. Only merges are a small exception here. If we look for the merge errors of graphics, we also take into account merges of graphic regions with other region types (e.g. separator).

**Image Area**

Image Width * Image Height

**Number of Black Pixel within the Image**

Number of black pixels within the black-and-white image.

**Overall Number of Regions**

Number of regions of the chosen level (region, text line, word or glyph) within the document layout.

**Number of Regions per Type**

Number of regions of for each region type (text, image, ...) within the document layout. This value is only available in region level and not in text line, word or glyph level.

**Overall Region Area**

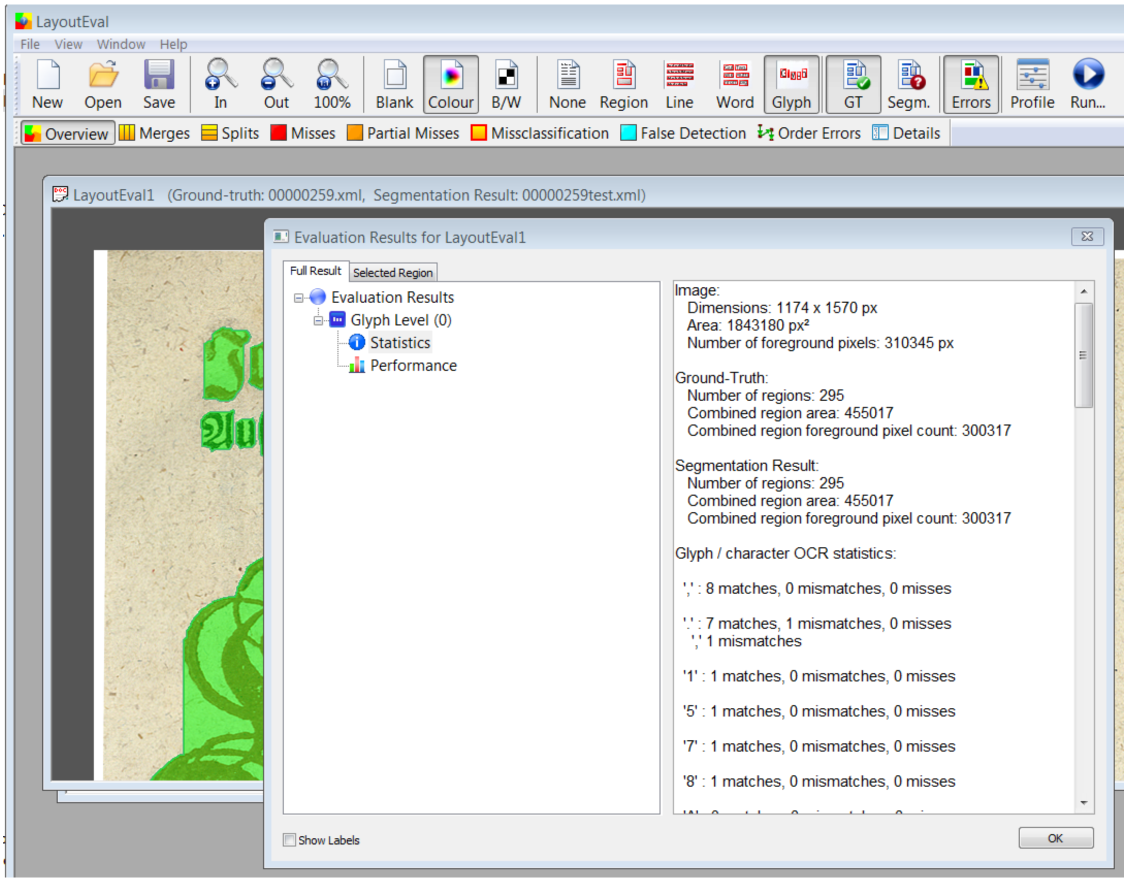
The combined area of all regions regarded for the 'Overall Number of Regions' value. Possible region overlaps are not left out. So if there are overlaps, some image parts are counted twice.

**Overall Number of Black Pixels in Regions**

The combined count of black pixels of all regions regarded for the 'Overall Number of Regions' value. Possible region overlaps are not left out. So if there are overlaps, some image parts are counted twice.

**Glyph / character OCR error details**

Each ground truth character is displayed in single quotation, followed by the number of correct matches, mismatch errors and miss errors. If the character was mismatched, details are shown below (one line for each type of mismatch + the count of how often the mismatch occurred).



LayoutEval1 (Ground-truth: 00000259.xml, Segmentation Result: 00000259test.xml)

Evaluation Results for LayoutEval1

```
Image:
  Dimensions: 1174 x 1570 px
  Area: 1843180 px²
  Number of foreground pixels: 310345 px

Ground-Truth:
  Number of regions: 295
  Combined region area: 455017
  Combined region foreground pixel count: 300317

Segmentation Result:
  Number of regions: 295
  Combined region area: 455017
  Combined region foreground pixel count: 300317

Glyph / character OCR statistics:

 ',' : 8 matches, 0 mismatches, 0 misses

 '.' : 7 matches, 1 mismatches, 0 misses
   ',' 1 mismatches

 '1' : 1 matches, 0 mismatches, 0 misses

 '5' : 1 matches, 0 mismatches, 0 misses

 '7' : 1 matches, 0 mismatches, 0 misses

 '8' : 1 matches, 0 mismatches, 0 misses
```

## Weighted Errors

For the weighted area and count the weights defined in the evaluation profile are being used. There are two types of weighted errors: the ‘Weighted Area’ and the ‘Weighted Count’. The weighted area is based on the assumption that bigger regions are more important than smaller ones. The error value is the region area (or the number of black pixels) multiplied with the weight. The weighted count only takes into account the error quantity. A misclassified region for instance is counted as one. A ground-truth region split into three regions is counted as 3. The count is then also multiplied with the weight.

### Overall Success Rates

The overall success rates combine all measure success rates to one number. There are two different types of overall success rates: the arithmetic mean and the harmonic mean. And for each type there are again two success rates: One including the weighted area success rates and the reading order and the other one with the weighted count success rates and the reading order.

The general formula for the weighted arithmetic mean is:

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i \cdot x_i}{\sum_{i=1}^{n} w_i}.$$

The general formula for the weighted harmonic mean is:

$$\frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} \frac{w_i}{x_i}}.$$

Where n is the number of values, $x_i$ are the values (in our case the success rates for the error types) and $w_i$ are the weights. For the reading order the weight is directly the one defined in the evaluation profile. The other weights are defined by:

$w_i = (5 * (1-x_i) + 1) / 6$

This highlights low error type success rates and diminishes high success rates, without erasing them completely. The influence of a partial success rate to the overall rate is somewhere between 1/6 and 1.

Note: If one of the error type success rates is zero, the harmonic mean is also zero.
The harmonic mean is always smaller than or equal to the arithmetic mean.

### Precision and Recall

Precision and recall are generally defined as follows:

*Precision* is the number of relevant documents retrieved by a search divided by the total number of relevant documents.

*Recall* is the number of relevant documents retrieved by a search divided by the total number of documents retrieved by that search.

In terms of document image evaluation this can be interpreted as follows (example for text regions):

Recall is the number of pixels within ground truth text regions that are also within a text region in the segmentation result divided by the overall number of pixels in ground truth text regions.

Precision is the number of pixels within ground truth text regions that are also within a text region in the segmentation result divided by the overall number of pixels in segmentation result text regions.

For the overall recall and precision we differentiate between strict and non-strict. For the strict recall and precision the region type must be matched correctly. That means a ground truth text region overlapped by a segmentation result image region does not count as recall. For non-strict recall and precision the region type doesn't matter. You could also say that strict means 'with classification' and non-strict means 'without classification'.

**F-Measure**

The F-Measure combines precision and recall to one quality measure. It is defined by:

F-Measure = 2 * precision * recall / (precision + recall)

**Region Count Deviation**

The region count deviation is simply the difference between the number of ground truth regions and the number of segmentation result regions.

absolute region count deviation = |#ground truth regions - #segmentation result regions|

relative region count deviation = absolute deviation / #ground truth regions

Note: If there are no ground truth regions, the relative value is the same as the absolute value.

# Creating an Evaluation Profile

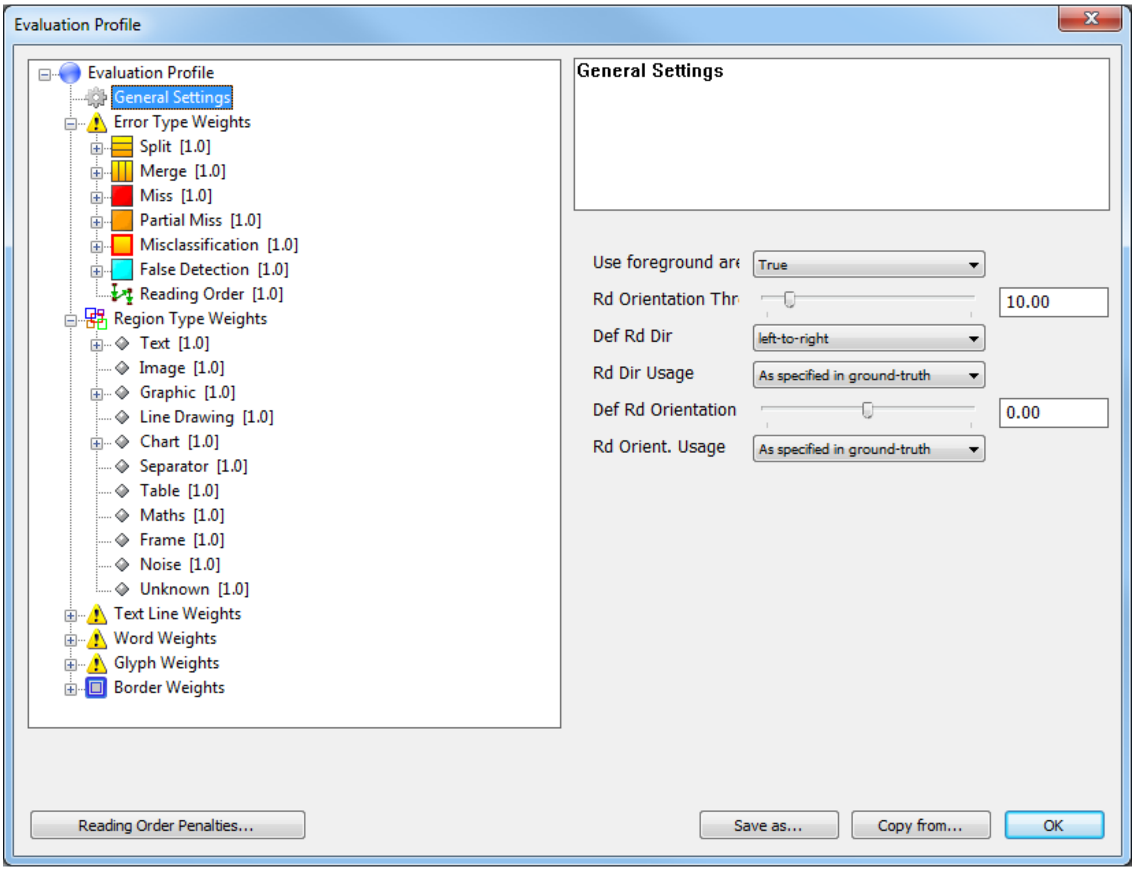
The evaluation profile is used to specify weights and other parameters to customise the layout evaluation. In some scenarios text regions may be important, in others it may be image regions. The weights can be adjusted from a general level (e.g. merge) down to the most detailed level (e.g. merge of text paragraph with text headline).
A weight can have values from 0.0 to 10.0, whereas 1.0 is the default. A value of 0.0 means that the region or error type is not regarded at all for the evaluation results. A value higher than 1.0 means that the region or measure (error type) is emphasized in comparison to other region or measures.
The profile is stored together with the evaluation results. So the weights that were used for the evaluation can always be examined.

The evaluation profile contains following weights and parameters:

- Error type weights stored hierarchical per error type (merge, split, ...), region type (text, image, ...) and for text/graphic/chart regions also subtype (paragraph, headline, ...). The weights for merge and split have an extra value for ‘allowable’ (see chapter on ‘Allowable Merges and Splits’).
- Region type weights stored per region type (text, image, ...)
- Reading order weight (influence of the reading order to the overall success rate)
- Reading order penalties (customisable penalty matrix that is used for reading order evaluation)
- General parameters
    - ‘Use Foreground Area’ – if TRUE, the number of black pixels is used for the error calculations instead of the polygon area
    - ‘Reading Orientation Threshold’ – Threshold of how much the reading orientation of two regions can differ to be allowable (see chapter on ‘Allowable Merges’)
    - ‘Default Reading Direction’ – Used for ‘Allowable Merge’ detection. See next parameter
    - ‘Reading Direction Usage’ – Can be either ‘Ground-truth’ – always use reading direction as specified in the ground-truth; ‘Default if not set in Ground-Truth’ – uses the default value if the reading direction isn’t defined for the regarded region; ‘Default’ – always uses the default value, regardless which value is defined for the regarded region
    - ‘Default Reading Orientation’ - Used for ‘Allowable Merge’ detection. See explanation above
    - ‘Reading Orientation Usage’ - Used for ‘Allowable Merge’ detection. See explanation above
    - ‘Default text type’ – Text region type that is to be used if not defined (use <undefined> to not use a default text type)
    - ‘Ignore embedded text misclassification’ – if TRUE, misclassification is not penalised if a chart, image, graphic, line drawing, or table region was detected as text region and the ‘embedded text’ attribute is set to TRUE in the ground truth region
    - ‘Evaluate nested regions’ – if TRUE, regions within regions (nested regions) are taken into account. Examples are: table cells, text within graphics or photographs, chart labels. Errors where nested regions are involved are weighted 50% (combined with any other weights)

To create a profile click on ‘Profile’ in the toolbar:

This will open the profile dialogue:



- Evaluation Profile
  - General Settings
  - Error Type Weights
    - Split [1.0]
    - Merge [1.0]
    - Miss [1.0]
    - Partial Miss [1.0]
    - Misclassification [1.0]
    - False Detection [1.0]
    - Reading Order [1.0]
  - Region Type Weights
    - Text [1.0]
    - Image [1.0]
    - Graphic [1.0]
    - Line Drawing [1.0]
    - Chart [1.0]
    - Separator [1.0]
    - Table [1.0]
    - Maths [1.0]
    - Frame [1.0]
    - Noise [1.0]
    - Unknown [1.0]
  - Text Line Weights
  - Word Weights
  - Glyph Weights
  - Border Weights

**General Settings**

Use foreground are: True

Rd Orientation Thr: 10.00

Def Rd Dir: left-to-right

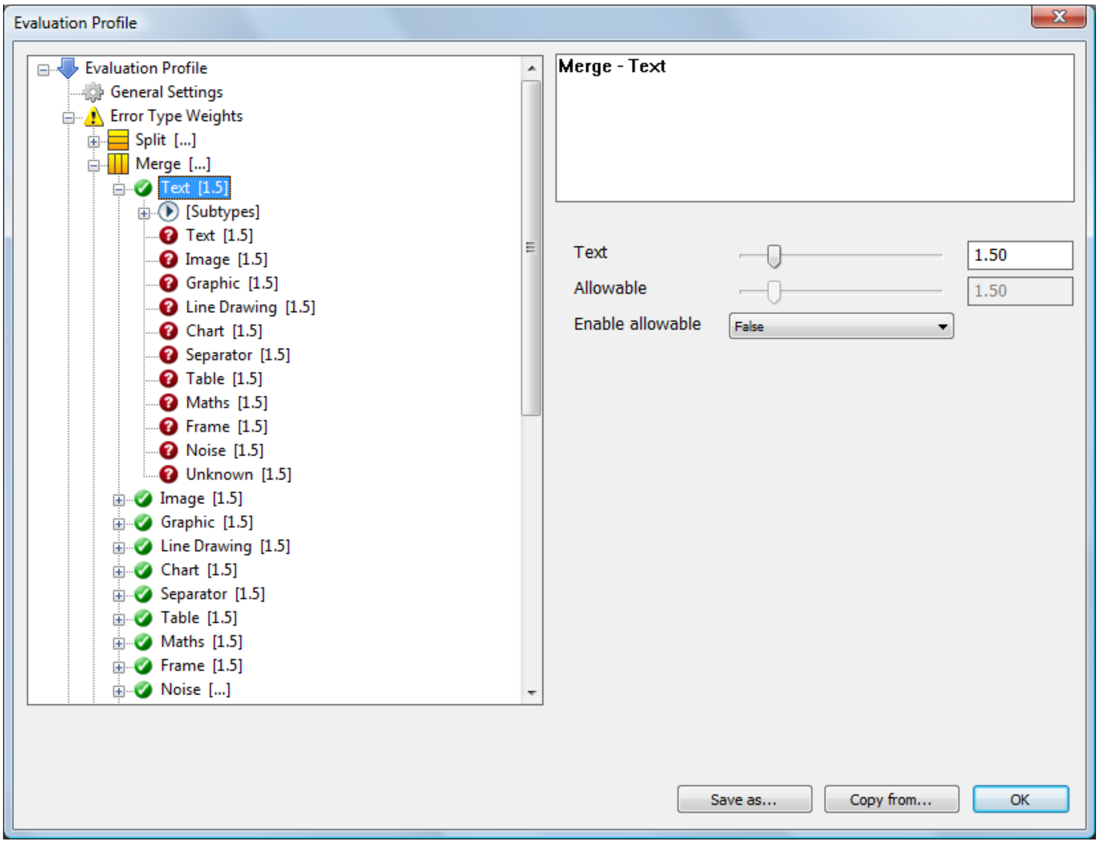
Rd Dir Usage: As specified in ground-truth

Def Rd Orientation: 0.00

Rd Orient. Usage: As specified in ground-truth

If you want to load an existing profile, click the button ‘Copy from...’ and select an .evx file. All settings will then be copied into the current profile.

To change a weight, select the desired weight node and adjust the slider on the right:



Text 1.50
Allowable 1.50
Enable allowable False

Note: Changing the weight within a parent node will change the weights of all child nodes. That way it is easy to change all merge weights at once and refine some sub weights separately afterwards.

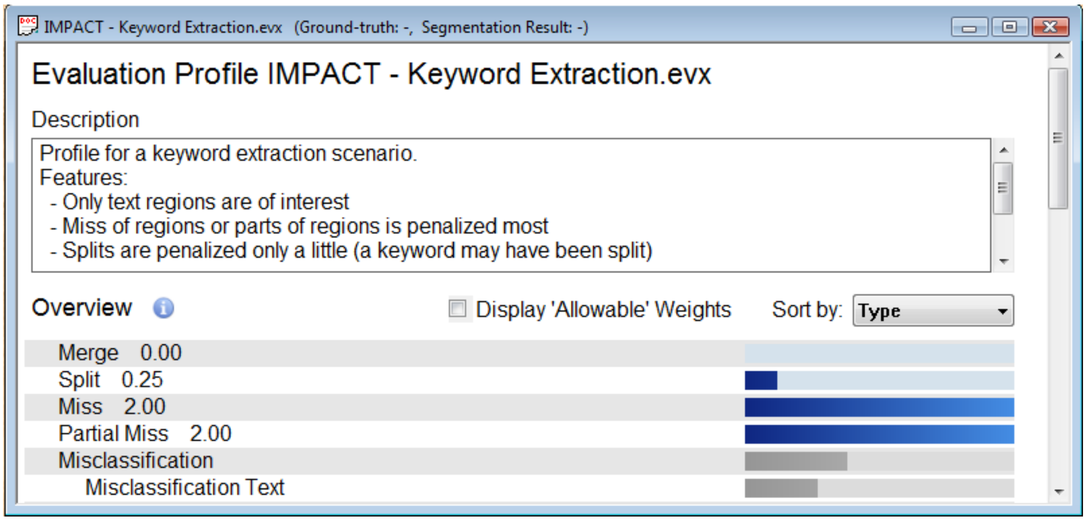
The allowable weights are available for merge and split only. To use the allowable, select ‘True’ from the drop-down box ‘Enable allowable’.

Finish with clicking ‘Save as...’ and following the save dialogue.
Note: Without saving the profile is not accessible for the evaluation and will be overwritten.

If a profile has been saved, it will automatically be selected for the next evaluation.

### ‘Profile only’ Mode

If a profile file is opened (an .evx file that contains only a profile and no evaluation data), the tool will automatically switch in the ‘profile only’ mode and shows an overview in the main document window area instead of a document image:



IMPACT - Keyword Extraction.evx (Ground-truth: -, Segmentation Result: -)

Evaluation Profile IMPACT - Keyword Extraction.evx

Description

Profile for a keyword extraction scenario.
Features:
- Only text regions are of interest
- Miss of regions or parts of regions is penalized most
- Splits are penalized only a little (a keyword may have been split)

Overview

□ Display 'Allowable' Weights

Sort by: Type

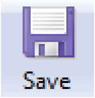
| | |
|---|---|
| Merge | 0.00 |
| Split | 0.25 |
| Miss | 2.00 |
| Partial Miss | 2.00 |
| Misclassification | |
| Misclassification Text | |

The overview includes an editable description field (uses the comments from the metadata) and a table of important weights. The weights can be sorted by type or relevance.

To create a profile file from scratch, you can click the 'New' toolbar button and press the 'Evaluation profile only' button in the dialogue. Then the profile dialogue will open automatically. Alternatively you can use the 'Create new profile' icon within the 'Welcome' window.

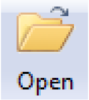
## Saving and Loading an Evaluation

It is possible to save an evaluation result. To do so click 'Save' in the toolbar and select a file name:



This will save everything: the used document files, the evaluation profile and the results.

To load an evaluation click 'Open':



This will open the 'Evaluation Files' dialogue and start loading the files. If all files were found, you can continue immediately, otherwise select the missing files manually.

# Credits

**PRImA Research Group**
University of Salford
United Kingdom

www.primaresearch.org

| | |
|---|---|
| Director of research group | *Apostolos Antonacopoulos* |
| Layout Evaluation project lead | *Christian Clausner* |
| Design | *Christian Clausner*<br>*Christos Papadopoulos*<br>*Stefan Pletschacher* |
| Development | *Christian Clausner* |
| Testing | *Christian Clausner*<br>*Stefan Pletschacher* |
| Documentation | *Christian Clausner* |