Aletheia User Guide

Version 4.1
Date: November 2019

www.primaresearch.org

Contents

Glossary

Border of a document image	Outer area of a document image that does not belong to the actual document.
Connected component	Here: Region of neighbouring foreground pixels.
Dewarping	Method to geometrically correct (straighten) images that appear warped in any way (e.g. due to aging, water damage, scanning process, etc.).
Document image	Digital copy of a document gained by scanning or photographing.
Glyph	Element of writing. Graphical representation of a letter.
Ground truth	‘Perfect’ page segmentation done by a human.

Isothetic polygons	A polygon is isothetic, if it consists only of horizontal and vertical lines.
Layer	Logical layout element to group regions. Allows regions to overlap each other without making the document layout invalid. Three typical layers for a document would be: back (e.g. for watermarks), middle (e.g. for text) and front (e.g. for stamps).
Layout region	Part of a document with specific properties (examples: text region, image region).
OCR	Optical character recognition is the mechanical or electronic translation of images of handwritten, typewritten or printed text (usually captured by a scanner) into machine-editable text.
Page Collection	A loose collection of document pages that can be saved and loaded.
Page Segmentation	Partitioning of a digital image into two or more regions.
PAGE XML format	XML format to store page segmentation results. The most recent schema file can be found here: http://schema.primaresearch.org/PAGE/gts/pagecontent/
Polyline	Connected series of line segments. Also known as polygonal chain or polygonal path.
Polygon	Closed series of connected line segments.
Print space	Main text body of a document without page numbers, marginalia, etc. The print space should cover all lines except: • page number (except together with page header) • marginalia • signature marks • preview words
Reading order	Order to read the text regions of a document.
Smearing	Filling the white space between pixels.

System Requirements and Installation

Minimum System Requirements

Operating System:	Windows 7 or higher
CPU:	2.5 GHz
RAM:	4 GB
Screen Resolution:	1280 * 800 (at 100% size setting in Windows)
Hard disk space:	2 GB

Recommended System Configuration

Operating System:	Windows 10
CPU:	3.0 GHz quad core
RAM:	16 GB
Screen Resolution:	1600 * 1200 (at 100% size setting in Windows)
Hard disk space:	250 GB

Internet Connection (required for automatic check for updates, see below)

Installation

There is no special installer for Aletheia. It is delivered as a bundle of files and is ready to use after copying it (or extracting it if it is compressed) to the local file system. However, if an older version of Aletheia was already in use on the same machine, a restart may be required for some changes to take effect (e.g. updated font).

Editions and Activation

Aletheia needs to be activated. If inactive, a Licence Management dialog with instructions will be shown automatically when starting Aletheia.

Go to www.primaresearch.org/tools/Aletheia/Editions to find out about available editions of Aletheia and how to get a licence (including free trial and free Aletheia Lite).

For more information see also the section on Licence Management.

Note:
In case a personal firewall is in use, an exception should be added to enable the online activation and allow Aletheia to check for updates.

Update

Aletheia automatically checks if a new version is available and shows an update hint and a link to the download page. You can also run a check manually (Help – Check for Updates).

About Aletheia

Aletheia is a document image analysis system. Its core function is to create and view *page segmentation* and *OCR ground truth*. The native storage format is *PAGE/*pagecontent (XML). Thus, Aletheia can also be used as a viewer for segmentation and OCR results produced by third party software supporting PAGE.

The following PAGE features are supported by this version of Aletheia:

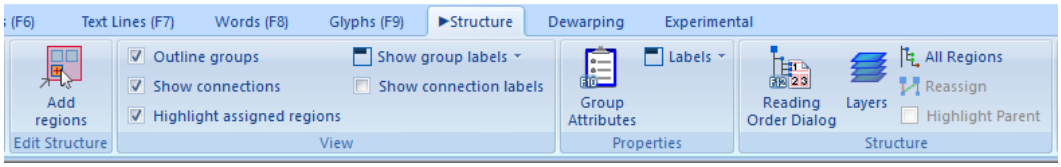
- Page elements on four levels (regions, text lines, words and glyphs), including
  - Polygonal outlines (semi-automated)
  - Element attributes (background colour, orientation angle, …)
  - Text content (Unicode, for text regions)
- Document border
- Document print space
- Reading order / page structure
- Region layers
- Metadata
- Dewarping data
- Page collections
- Performance evaluation / segmentation result quality measurement

See the document ‘Aletheia Introduction’ for more information.

New in Version 4

Significant changes in Aletheia 4 compared to its predecessor version 3.4 are highlighted using text boxes marked with *3.4↔4.0.*

Page structure:



The reading order now has a dedicated toolbar tab called “Structure”. In this view, reading order groups are now outlined (as rectangles) as image overlay. The reading order dialog works as before.

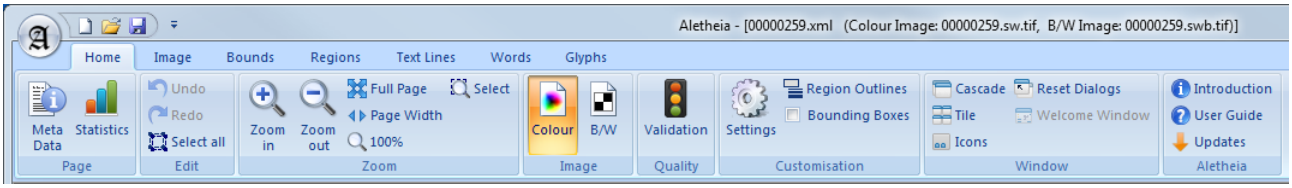
Tesseract OCR engine update:

Aletheia now comes with Tesseract 4.0. The previous version (3.04) is still available and can be selected in the settings.

Using Aletheia

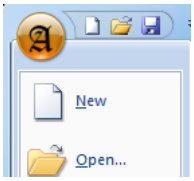
Toolbars

Aletheia uses the Microsoft Office style ribbon toolbar:



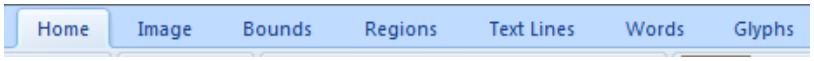
It consists of:

- An Aletheia button and file menu:

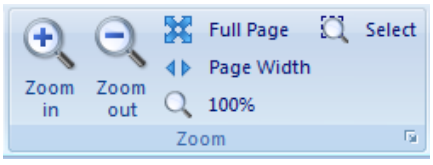


- A quick access toolbar (configurable):

- Categories (tabs):

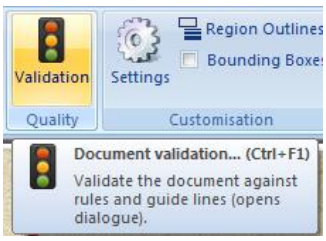


- Panels (sections within a category):



↳ Some panels have a launch button that opens a popup menu with all the actions of the panel. This can be useful if the labels of the icons are hidden due to lack of space.

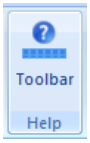
- Extended tooltips:



The toolbar adjusts itself to the size of the window. If the space is not sufficient big icons will be changed to small icons, text labels will be removed and finally whole panels will be replaced by drop-down menus.

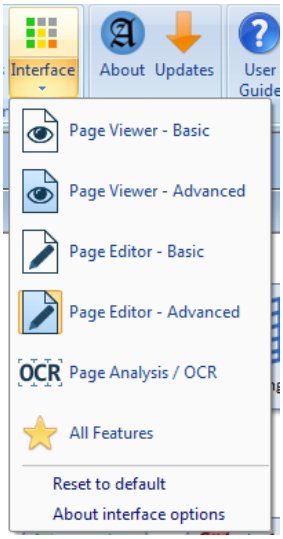
Context sensitive help:

Look for following button to open a help file on the current toolbar tab:



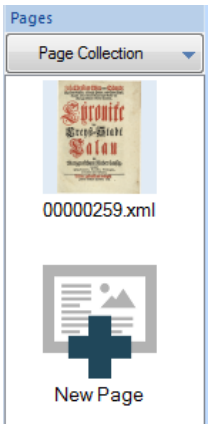
Changing the toolbar:

It is possible to switch the toolbar using a range of pre-defined layouts.
Use the “Interface” menu to select another toolbar layout:



Have a look at the dedicated documentation to learn more (“About interface options” in the menu).

On the left side there is an additional tool pane for Page Collections.



Creating, Opening, Saving, and Exporting Documents

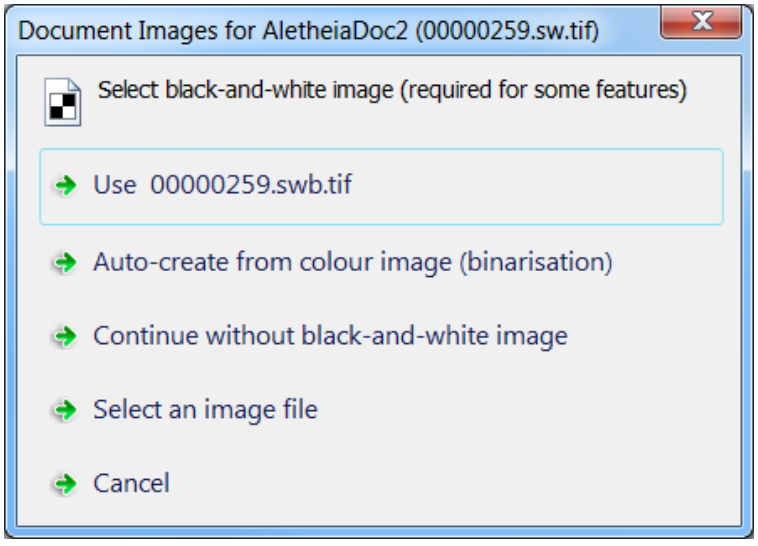
Creating a new Document

To create a new document:

- Click the New button of the quick access toolbar or use the corresponding entry in the Aletheia menu (keyboard shortcut Ctrl+N)

- Select a document image file within the dialog that opens
  - Supported formats: TIFF, PNG, JPEG, JPEG2000 (limited), PDF (limited)
- If available, select the corresponding black-and-white or colour image, otherwise click “Auto-create …” or “Continue without ...”

  Since Aletheia 3.3, the dialog for choosing a secondary image is only shown on demand (when clicking the message or when using a tool that requires a black-and-white image).



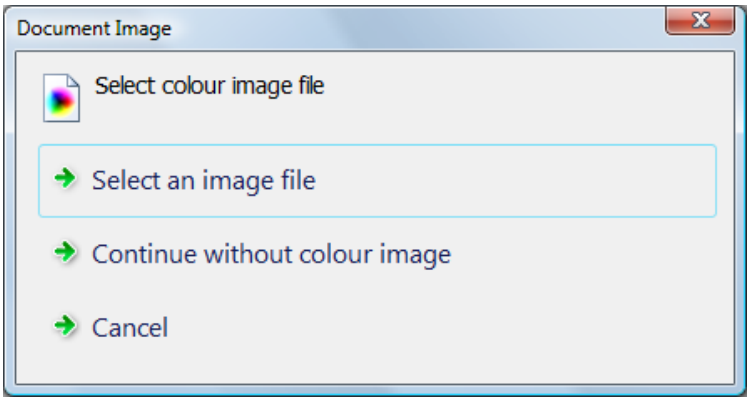
Note:
Providing a black-and-white image is optional. However, without black-and-white image some features of Aletheia will not be accessible.

Opening an Existing Document

To open an existing document:

- Click the Open button of the quick access toolbar or use the corresponding entry in the Aletheia menu (keyboard shortcut Ctrl+O)

- Select an XML file in PAGE format within the dialog that opens
- (Select the associated document image(s) using the dialog)



The XML file contains the file name of the colour image. If the image is within the same folder it will be loaded automatically.

Note:
Providing a black-and-white image is optional. However, without black-and-white image some features of Aletheia will not be accessible.

Supported XML Formats

PAGE XML

Aletheia's native format is defined as part of PAGE (Page Analysis and Ground truth Elements). Documents are always saved in the latest version of PAGE XML. It is however possible to open documents that are stored using older versions of PAGE.

FineReader XML

Aletheia can read Abbyy FineReader XML files (schema versions 6.1 and 10.1).
Following restrictions apply:

- Experimental feature (correctness of layout cannot be guaranteed)
- Internet connection required (to access the ABBYY XML schema)

- Read-only (documents are always saved in PAGE XML format)
- Multiple pages are not supported (only the first page will be opened)

ALTO XML

Aletheia can read ALTO XML files (schema versions 1.4, 2.0, 2.1, 3.0, and 4.0).
Following restrictions apply:

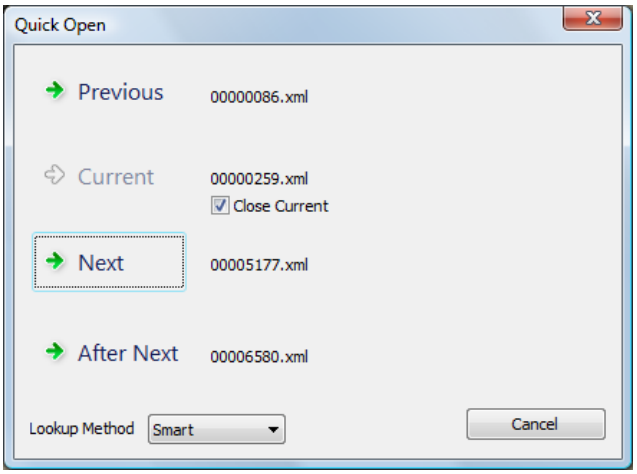
- Experimental feature (correctness of layout cannot be guaranteed)
- Read-only (documents are always saved in PAGE XML format)
- Multiple pages are not supported (only the first page will be opened)

Quick Open Function

The 'Quick Open' feature speeds up workflows with multiple documents involved.

To use this function:

- Click 'Quick Open' in the Aletheia menu (keyboard shortcut Ctrl+Q)
- Select a document to open ('Previous', 'Next' or 'After Next')



The dialog proposes the immediately preceding and succeeding files in the sequence of documents. By default the current document will be closed when opening another one. To change this behaviour, unselect the check box 'Close Current'.

There are different methods how the proposed documents are determined. If the current method doesn't show the expected files, try the other methods by selecting them from the drop down box at the bottom.

Note:
The Quick Open dialog is only available if at least one document is already open. The location of the current document is used to determine the documents proposed in the dialog.

Saving a Document

Different page operations affect different files. A “Save all changes” option and context-sensitive save are available.

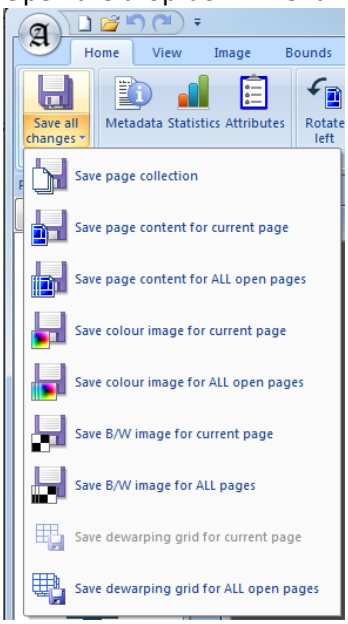
To **save all changes** in all open pages and the page collection:

- Click on “Save all changes” on the toolbar

- Choose locations if prompted

To **save individual changes** for the current page or all open pages:

- Open the drop-down menu from the “Save all changes” button



- Select what to save

To **save what you are working on** (image, page content, or dewarping grid):

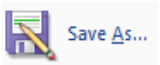
- Click on “Save”

OR

- Press CTRL + S

To save the page content XML using a different name:

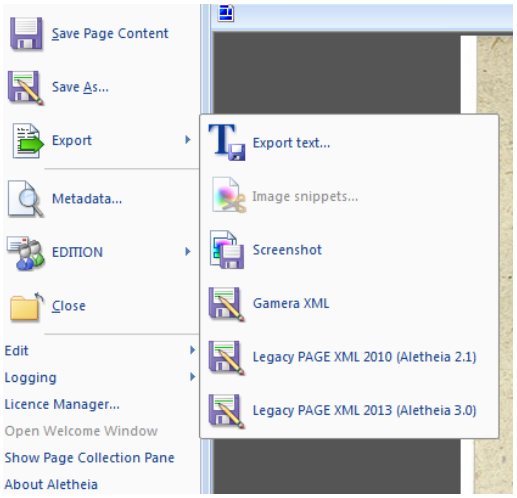
- Use the 'Save as' entry in the Aletheia menu



Exporting a Document or Parts of a Document

To export as Gamera classifier XML file:

- Select “Export” – “Gamera XML” from the Aletheia menu



- Choose a file location and click “Save”

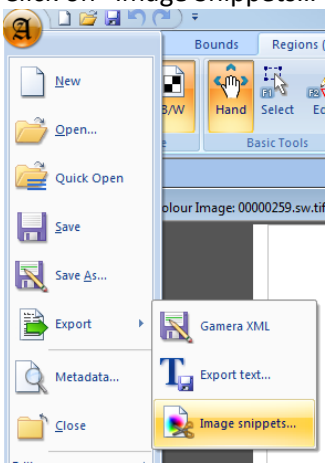
Note: The export requires a look-up table file for character class names (located in [Aletheia]\bin\data\Gamera). Modify the look-up table using a default XML editor as required.

To export the text content:

- See the ‘Text Export’ section in chapter ‘Page Object Properties and Text Content’

To export image snippets (cut-outs of the full document image):

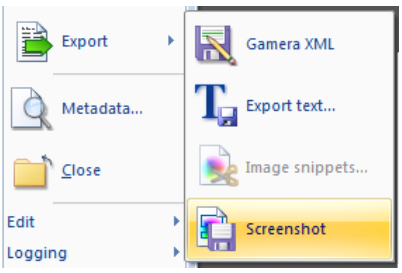
- Select the page objects (regions, text lines, words, or glyphs) for which to cut out image snippets
- Optional: Switch to the image you want to use (colour or black-and-white image)
- Click on “Image Snippets…” in the Aletheia sub-menu “Export”



- Select a target folder
  - The snippets are saved as new image files within this folder, using the filename of the original image as base and the object ID as ending

To export a screenshot (whole image with overlays):

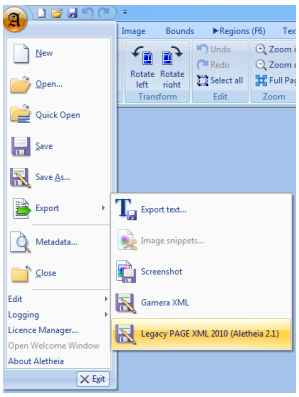
- Optional: Zoom out to create a smaller image (exporting with zoom over 100% is not possible)
- Click on “Screenshot” in the Aletheia sub-menu “Export”



- Specify target image file (PNG)

To export in legacy PAGE 2010/2013/2016/2017 XML formats:

- Click on one of the “Legacy PAGE XML 20xx (Aletheia x.y)” in the Aletheia sub-menu “Export”
- Choose file location and name
- Click on save



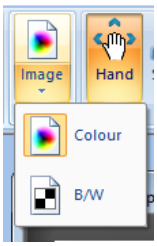
Note: Some information might be lost when saving in an older format.

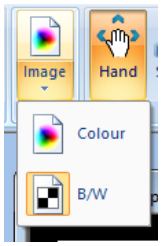
Viewing Documents

Selecting the Document Image

To select type document image to display (colour/greyscale or black-and-white (bitonal)):

- Click one of the icons within the image panel of the toolbar (keyboard shortcut TAB – toggles between colour and black-and-white image)





To fade the document image to transparency (make it fainter, to highlight graphical overlays):

- Switch to the 'View' toolbar tab
- Use the 'Transparency' slider in the 'Image' panel

Zooming

Basic zoom functions (available in every tab of the toolbar):

Zoom in (shortcut + on num pad or Ctrl + mouse wheel)

Zoom out (shortcut – on num pad or Ctrl + mouse wheel)

Full Page (shortcut / on num pad)
Fits the full document into the window.

Extended zoom functions (only available in the 'Home' toolbar section):

Page Width (shortcut decimal point on num pad)
Fits the document to the width of the window.

100% (shortcut * on num pad)
Resets the zoom to 100%.

Select
Tool to select an area to zoom in to.

Thumbnail
Shows or hides a thumbnail of the document image with a highlight of the current view position.

If the document doesn't fit to the screen, scroll bars will appear to the right and/or the bottom of the viewing area.

It is also possible to set a default zoom type in the user settings. This zoom type will then be used when a new document is created or a document is opened. Three types are available:

- '100%'
- 'Fit to window'
- 'Last used zoom level'

The Hand Tool

The hand tool provides a convenient way for document scrolling (panning). It can be activated using the 'Hand' toolbar button (keyboard shortcut Space). Mouse click and drag scrolls the view in any direction. As of Aletheia 3.1, the "Select" tool provides the same panning functionality when using the right mouse button.

The keyboard shortcut (Space) can be used to toggle between the previously used tool and the hand tool.

The hand tool can also be used to select regions (see section Selecting Regions).

The hand tool can be used to resize a page layout object.
To resize a region:

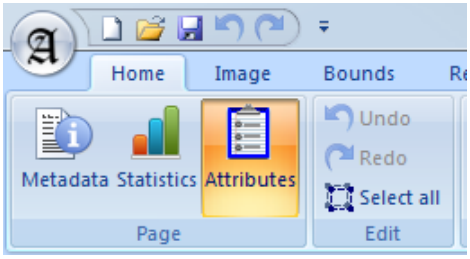
- Select a page layout object (e.g. a region)
- Place the mouse cursor over the selection rectangle and press SHIFT (the cursor changes)
- Click and drag the mouse

A cross-hair aid at the current mouse position can be enabled at any time by pressing CTRL and SHIFT while moving the mouse cursor.

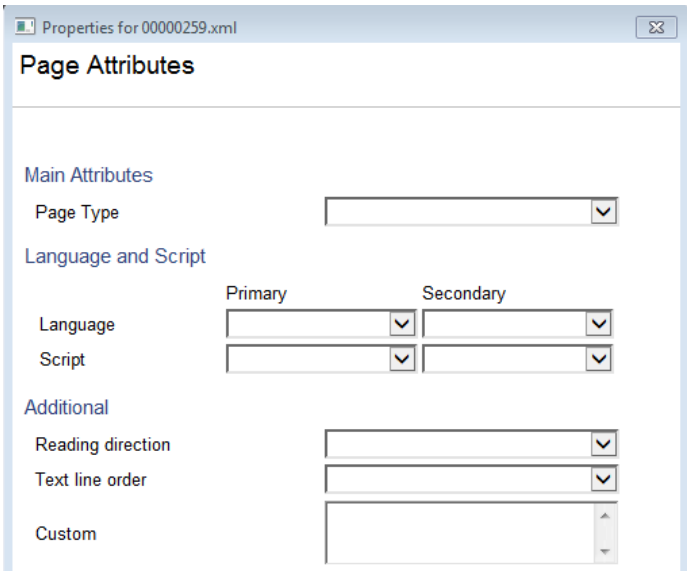
Page Attributes

To view and modify page attributes:

- Switch to category ‘Home’ within the toolbar and click on ‘Attributes’ to open the dialog



- Modify the fields as required and click ‘Ok’



Custom Attributes

It is possible to add more page attributes.

To add user-defined page attributes:

- Click on the “+” button at the bottom

- In the dialog:
  - Add attributes:
    - Click on Add and enter a name
    - Select a type
      - Text (e.g. “Water damage”)
      - Integer number (e.g. “23”)
      - Decimal number (e.g. “8.5”)
      - Boolean (“true” or “false”)
  - Edit and attribute:
    - Select and attribute and select a new type or enter a new value
  - Remove an existing attribute:
    - Select an attribute and click on Remove
  - Save or load an attribute profile
    - Expand the “Import attributes” drop-down list

- Select “Save…” to save your current attributes as profile for reuse
- Select a saved profile to add all attributes of the profile

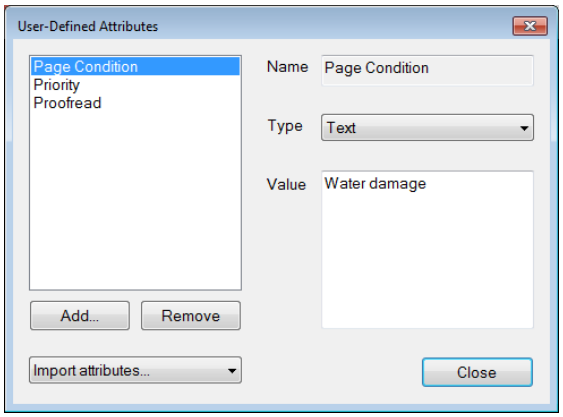
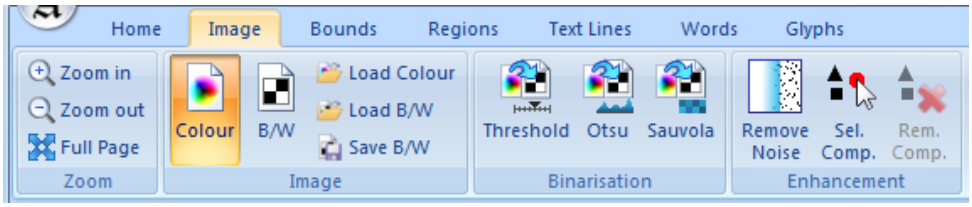


Image Tools

To use the image tools:

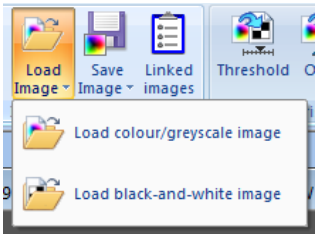
- Switch to category ‘Image’ within the toolbar



Loading and Saving Images

To load or reload a document image (colour, grey scale or black-and-white):

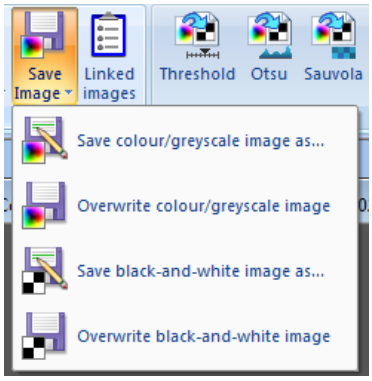
- Click the appropriate button in Image panel of the toolbar



- Select an image file using the dialog

To save an image:

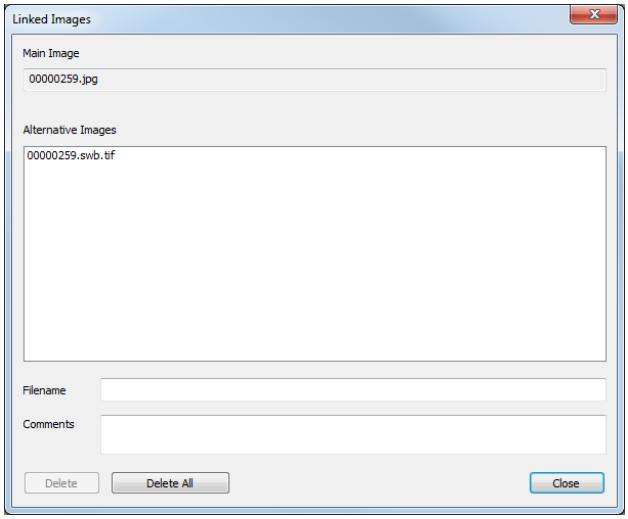
- Click the appropriate button in Image panel of the toolbar
  - Use “Save … as” to save a copy of the image
  - Use “Overwrite …” to replace the current image



- Select a destination and filename in the 'Save As' dialog

Aletheia stores the filenames of images linked to the current document.
To manage the linked images:

- Click on "Linked Images" in the Image panel of the toolbar
- A dialog opens:



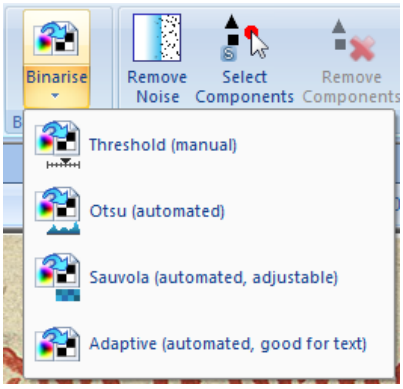
- Select an image in the listbox to view and/or change details of a linked alternative image

Binarisation

A colour or greyscale document image can be used to create a corresponding black-and-white image (bitonal). This is called binarisation. Some features of Aletheia require a black-and-white image to function.

To binarise the colour or greyscale image:

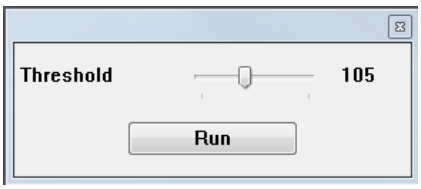
- Select one of the available methods in the 'Binarisation' panel of the toolbar



- (Adjust method parameter(s) if necessary)
- (Click 'Run')

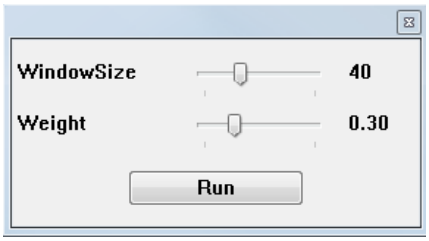
Available methods:

- Threshold: Simple method using a manually selected threshold to map grey scale values below the threshold to black and values above the threshold to white (colour images are converted to greyscale format internally)



- Otsu method: Similar to the threshold method, but the threshold is determined automatically.

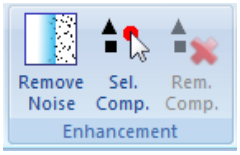
- Sauvola method: Adaptive approach subdividing the image in smaller areas (windows) and applying local thresholds. Especially suited for documents with uneven backgrounds.



- Adaptive method: Method by Gatos et al. (2005) "Adaptive degraded document image binarization". Suited for textual content.

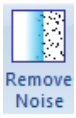
Noise Removal

Binarisation often leads to some noise in the resulting black-and-white image. This can complicate the ground truthing considerably. Thus, Aletheia offers several noise removal tools:

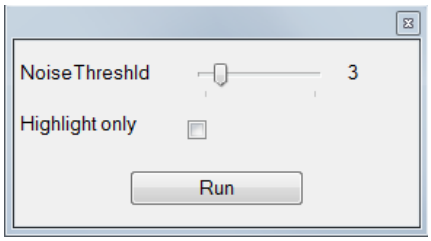


To automatically remove or highlight small black spots (‘pepper’ noise):

- Click ‘Remove Noise’



- Select a threshold for the size of the black spots (components)
- Optional: Tick ‘Highlight only’ to only select noise components (for inspection and manual removal)
- Press ‘Run’



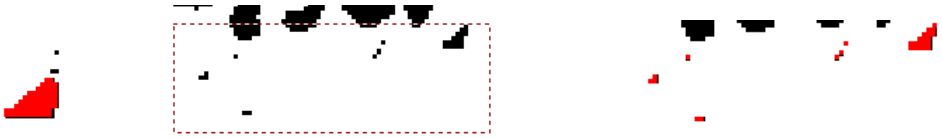
Examples:

‘Highlight only’ (the i-dot can be deselected before removing the noise components):

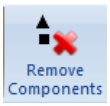
To manually remove selected black objects (connected components):

- Click ‘Select Components’

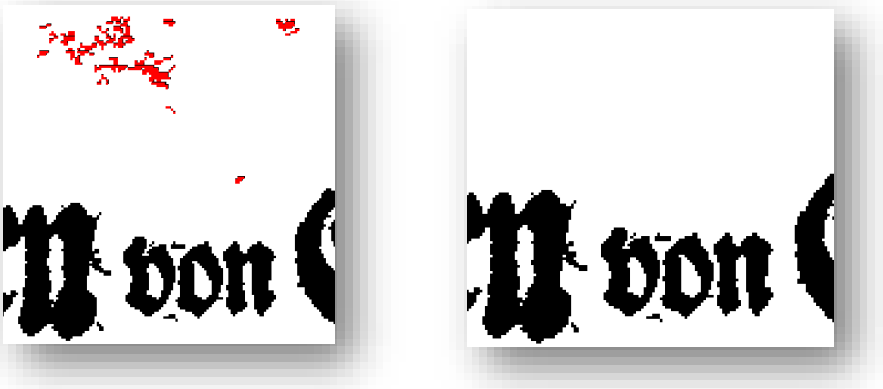
- Click on a black object (connected component) or drag a rectangle around multiple black objects. This selects one or multiple components (highlighted in red).



- (Repeat the previous step while pressing Ctrl or Shift to add or remove more components to the selection)
- Click ‘Remove Components’ (keyboard shortcut Del)



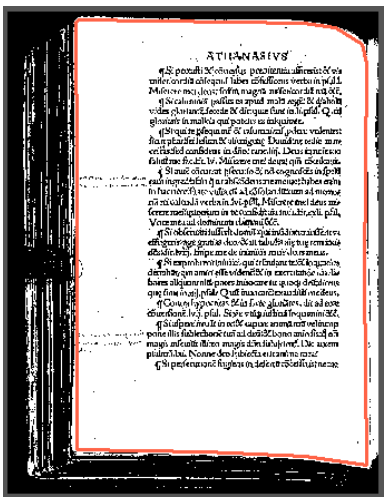
Example:

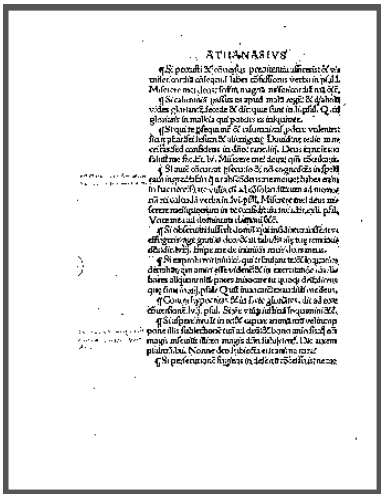


To remove all black pixels that lie outside the defined document border:

- Mark the document border (see next chapter)
- Click on ‘Erase Border’ in the ‘Enhancement’ toolbar panel

Example:



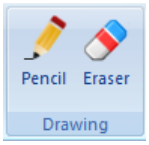


Drawing Tools

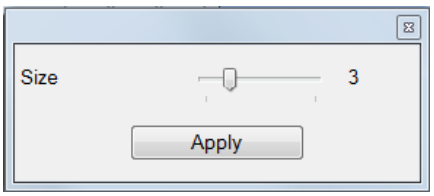
The black-and-white image can be modified on pixel level.

To add or remove black pixels from the black-and-white image:

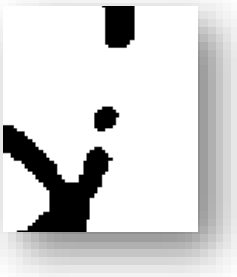
- Activate ‘Pencil’ or ‘Eraser’ from the toolbar panel called ‘Drawing’

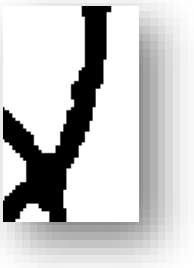


- Optional: Adjust the tool size using the slider



- Use the mouse (left button) for drawing (free-hand)





- Click on ‘Apply’ to confirm the changes

  OR

- Switch back to the default tool (ESC) to discard the changes

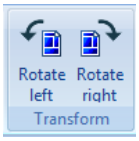
Note: It is possible to switch between pencil and eraser without applying the changes.

Rotation

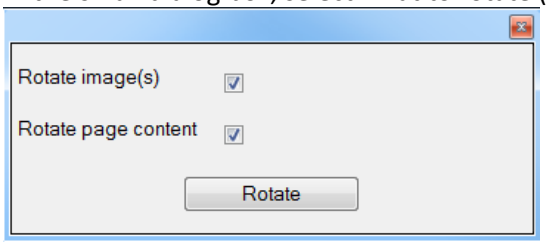
It is possible to rotate the document image(s) by 90 degrees.

To rotate:

- Activate one of the 'Rotate' tools (clockwise or counter clockwise)



- In the small dialog box, select what to rotate (image(s) and or page content (regions etc.)



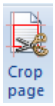
- Click on 'Rotate'

Cropping

It is possible to crop the document image(s) and the page layout to remove unwanted areas.

To crop:

- Activate the cropping tool

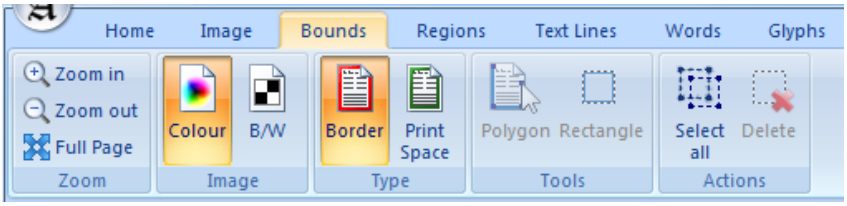


- Draw a rectangle demarking the new dimensions of the page
- (Save page content and images as required)

Document Bounds (Border and Print Space)

To view and edit document bounds:

- Activate the category ‘Bounds’ within the toolbar



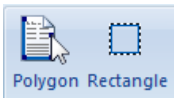
Marking the Document Border

The document border is the outer area of the image that does not belong to the actual document.

To mark the border:

- (Optional: Delete the existing border using the delete key or the toolbar icon)
- Select ‘Border’ in toolbar panel labelled ‘Type’ (keyboard shortcut F4)

- Click ‘Polygon’ or ‘Rectangle’ and mark the border



For more information see the chapters 'Drawing Polygons' and 'Drawing Rectangles'.

Example:

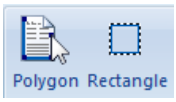
Marking the Print Space

The print space is the area containing only the main text body of the page without page number, marginalia etc. It is identical for all pages of a book.

To mark the print space:

- (Optional: Delete the existing print space using the delete key or the toolbar icon)
- Select ‘Print Space’ in toolbar panel labelled ‘Type’ (keyboard shortcut F5)

- Click ‘Polygon’ or ‘Rectangle’ and mark the print space



For more information see the chapters 'Drawing Polygons' and 'Drawing Rectangles'.

Example:


innerhalb des Felsens reitau, durch weite Hallen und Gallerien, die hier ausgesprengt und ausgemeisselt sind, und sieht furchtbare Batterien aufgestellt. Der 26. Mai sollte uns ein grossartiges Schauspiel bieten. An diesem Tage, der Feier des Geburtsfestes der Königin von England, war die ganze Besatzung auf der Ebene unter dem Felsen zur Parade ausgerückt, die Felsbatterien spielten und als von der obersten Spitze bis herab nach einander die Kanonen blitzten und donnerten, wiederhallte der ganze Fels und der Pulverdampf drang allenthalben qualmend hervor, wie Rauch aus einem Vulcane. Was mochten sich die vielen spanischen Officiere gedacht haben, welche in der Suite des englischen Gouverneurs waren, als so der ganze Fels Feuer und Rauch gegen Spanien auswarf? Wir hatten es gut getroffen, denn diese Feier verschaffte uns nicht blos das interessante militärische Schauspiel, sondern Abends auch den Genuss eines glänzenden Balles, den der Gouverneur Sir James Fargusson gab — glänzend durch die Schönheit der spanischen und englischen Damen, glänzend durch die mannigfaltigen Uniformen, die vertreten waren: englisch, schottisch, spanisch, österreichisch und französisch.

Auch dem Naturforscher bietet der Fels von Gibraltar manches Interessante. Er besteht aus einem stark zerklüfteten, lichten jurassischen Kalk mit mehreren Höhlen. Die schönste und grösste dieser Höhlen ist die St. Michaelsgrotte an der Südwestseite, eine hübsche Tropfsteingrotte, für deren Zugänglichkeit aber bis jetzt wenig gesorgt ist. Die Zoologen machten an dem Kalkfelsen eine reiche Ausbeute an Schnecken. Unten am Fels sind an der Westseite hübsche Gartenanlagen, an der Ostseite ist er ganz kahl. Der Gipfel aber ist bedeckt von dichtem Buschwerk, einer kleinen Fächerpalme (Chamaerops humilis) und den gelben Blüthenbüschen des Pfriemenstrauches. Am südlichsten höchsten Gipfel (1400 Fuss) sollen noch einige Affen vorkommen, die hier ihr Gaudenbrot bekommen, weil sie einst wie die capitolinischen Gänse bei einer Belagerung durch ihre Wachsamkeit sich auszeichneten. Da der Fels so

Automated Border Detection

To auto-detect the document border (experimental):

- Switch to the image that is to be used for the detection (colour or black-and-white)
- Click on ‘Detect Border’

Marking Layout Regions

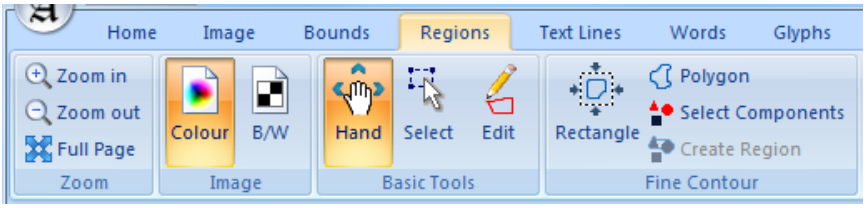
Layout regions are the result of the page segmentation and classification process. The PAGE format contains several region types such as:

- Text
- Image
- Table
- Separator
- ...

Text lines, words and glyphs are logical sub-regions of text regions and will be handled in separate sections.

To create, view and edit layout regions:

- Activate the category 'Regions' within the toolbar (keyboard shortcut F6)



Assisted Region Creation

Aletheia provides several tools to assist the creation of regions by automatically finding the contour. The user only has to roughly outline the area of the region by drawing a rectangle, a polygon or by selecting black components.
Some of the tools can also be used to recalculate the outline of an existing region.

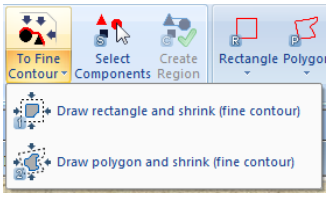
> **Note:**
> All tools for assisted region creation require a black-and-white document image.

Tools for Fine Contour

Rectangle and Polygon

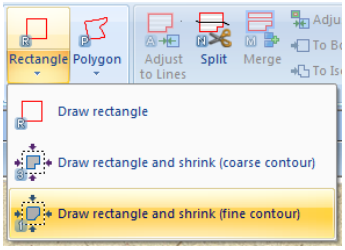
To create a region with contour detection:

- Activate ‘Rectangle’ or ‘Polygon’ from the ‘Fine Contour’ drop-down menu of the toolbar panel called ‘Auto Contour’ (keyboard shortcuts 1 and 2)

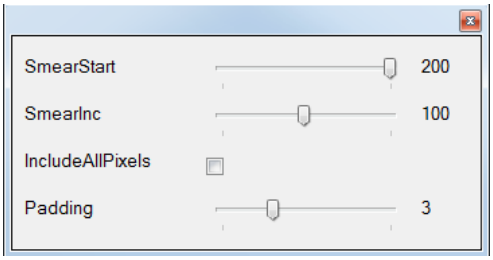


OR

- Select the tools from the respective ‘Rectangle’ or ‘Polygon’ drop-down menu

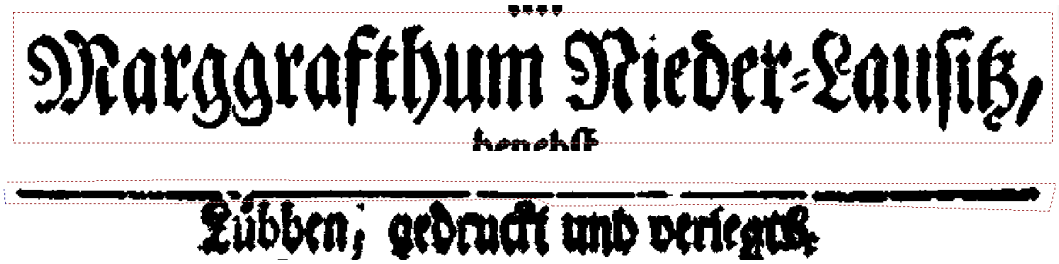


- Optional for advanced users: Adjust the parameters for the tool



  - Initial Smearing Value and Smearing Increment: The tool uses a *smearing* algorithm to connect all included connected components to one single component. The algorithm starts with an initial smearing threshold and increases this threshold until only one component remains.
  - Include all pixels: If selected, all black pixels inside the specified shrinking area are regarded during the shrinking process, even if they belong to a black component that is partly outside the area. If not selected, black components that are partly outside the shrinking area are disregarded. Examples:

    - Marked area



    - Include all pixels disabled

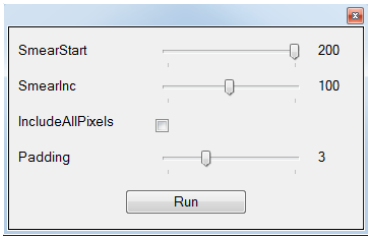
    - Include all pixels enabled

- Padding: Additional padding in pixels to be added around the region content

- Roughly specify the outline of the region using the selected drawing tool (see the sections 'Drawing Polygons' and 'Drawing Rectangles').

To recalculate the outline of an existing region:

- Select the region (see section ‘Selecting Regions’)
- Activate any Fine Contour tool
- Click ‘Run’ in the parameter window



Example:

Selecting Components

Region outlines can also be defined by selecting all connected components (black objects) belonging to the region.

To create a region with contour detection around a selection of components:

- Activate 'Select Components' (keyboard shortcut S)

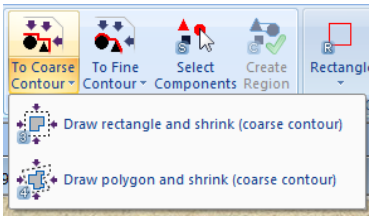
- Select components by
  - Left click on a component OR
  - Dragging a rectangle around components
- (Selected components are marked red)
- Add other components by clicking left or dragging a rectangle with the 'Ctrl' key pressed. Use ‘Shift’ and the mouse to toggle the current selection.
- Click 'Create Region' (keyboard shortcut C)

Example:

Tools for Coarse Contour

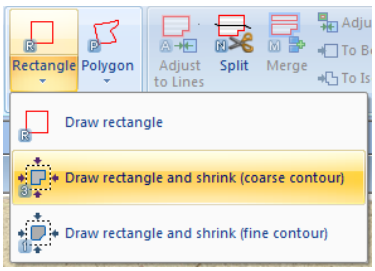
To create a region with coarse contour detection:

- Activate ‘Rectangle’ or ‘Polygon’ the ‘Coarse Contour’ drop-down menu of the toolbar panel called ‘Auto Contour’ (keyboard shortcuts 3 and 4)

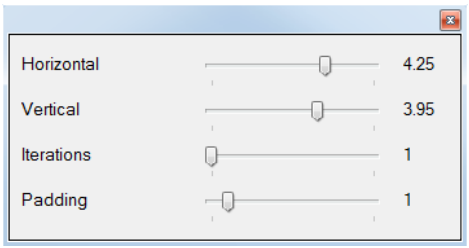


OR

- Select the tools from the respective ‘Rectangle’ or ‘Polygon’ drop-down menu



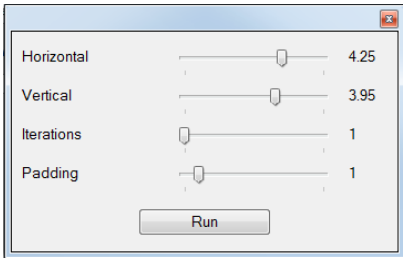
- Optional for advanced users: Adjust the parameters for the tool:
  - Horizontal, Vertical: ‘Simplicity’ of the resulting polygon (low values shrink deeper and high values shrink less)
  - Iterations: Count of shrinking iterations (by using more iterations it is shrunk deeper inside the regions)
  - Padding: Additional padding in pixels to be added around the region content



- Roughly specify the outline of the region using the selected drawing tool (see sections 'Drawing Polygons' and 'Drawing Rectangles').

To recalculate the outline of an existing region:

- Select the region (see section ‘Selecting Regions’)
- Activate any Coarse Contour tool
- Click ‘Run’ in the parameter window

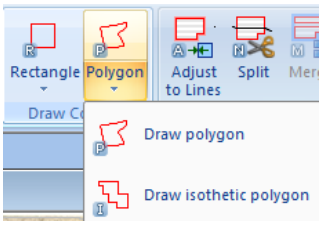


Example:

Defining Region Outlines Manually

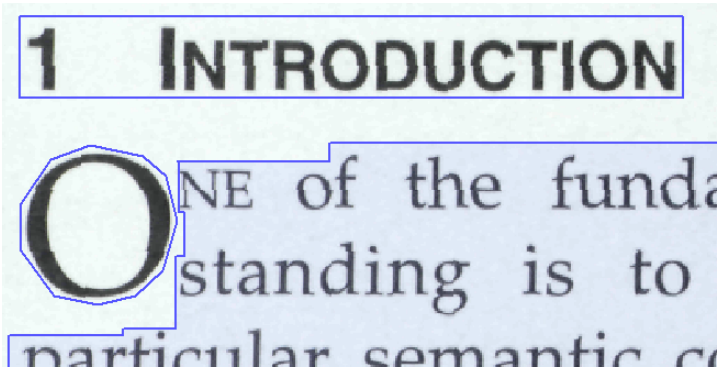
To create a region by drawing its outline:

- Activate Rectangle, Polygon or Isothetic polygon from the toolbar panel called ‘Draw Contour’ (keyboard shortcuts R, P and I)



- Draw the intended outline (see sections 'Drawing Polygons' and 'Drawing Rectangles')

Examples (top: rectangle; bottom left: polygon; bottom right: isothetic polygon):



The rectangle tool can also be used to create a region around a connected component. To do so:

- Switch to the black-and-white image or create one
- Activate the rectangle tool
- Right-click on a black object

Correcting Regions

This section explains tools to adjust and correct exiting layout regions. Note that some tools for contour detection can also be used to recalculate the outline of existing region (see the previous section).

Adjusting Region Outlines to Child Text Objects

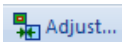
To adjust the outline of a text region to its text line objects:

- Select the region (see section ‘Selecting Regions’)
- Click ‘Adjust to lines’ in the toolbar panel called ‘Correction’ (keyboard shortcut A)

Example:

To adjust the outline of a text region to its child glyphs, words and text lines:

- Select the region (see section 'Selecting Regions')
- Click 'Adjust...' in the toolbar panel called 'Correction'



Splitting Regions

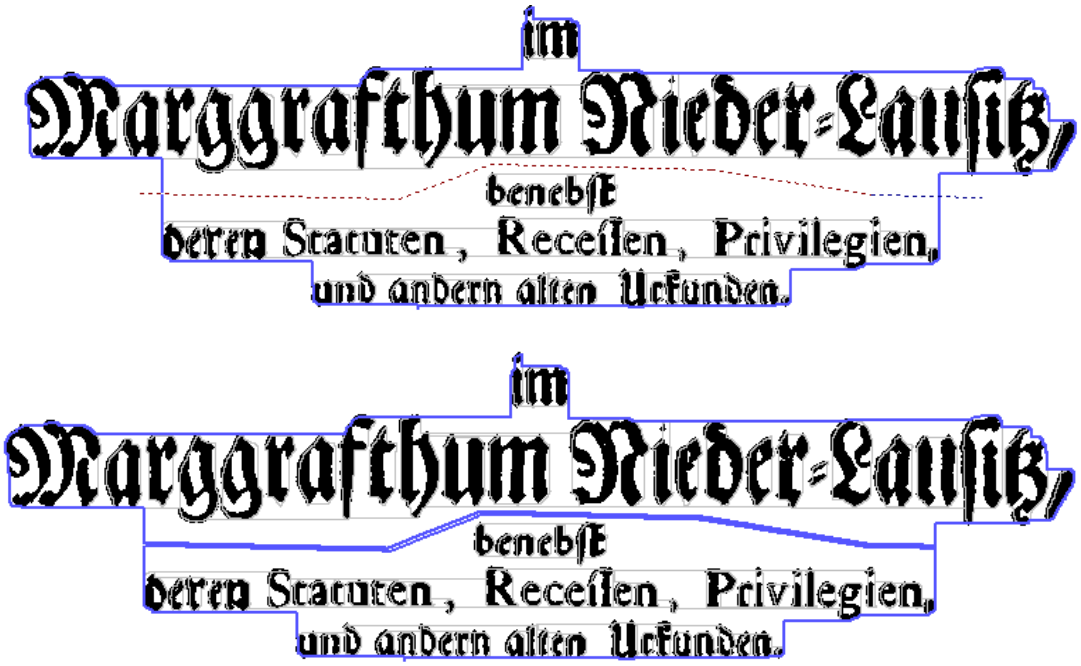
To split a region into new regions:

- Activate 'Split' in the toolbar panel called 'Correction' (keyboard shortcut N)

- Draw a split line across the region (starting and ending outside the region) (see section 'Drawing polygons and polylines')
- (If the region contained text, revise and correct the text of the new regions)

Important:
The text content of the split region is fully copied to each new region. Hence the text always has to be revised after a split.

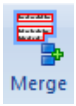
Example:

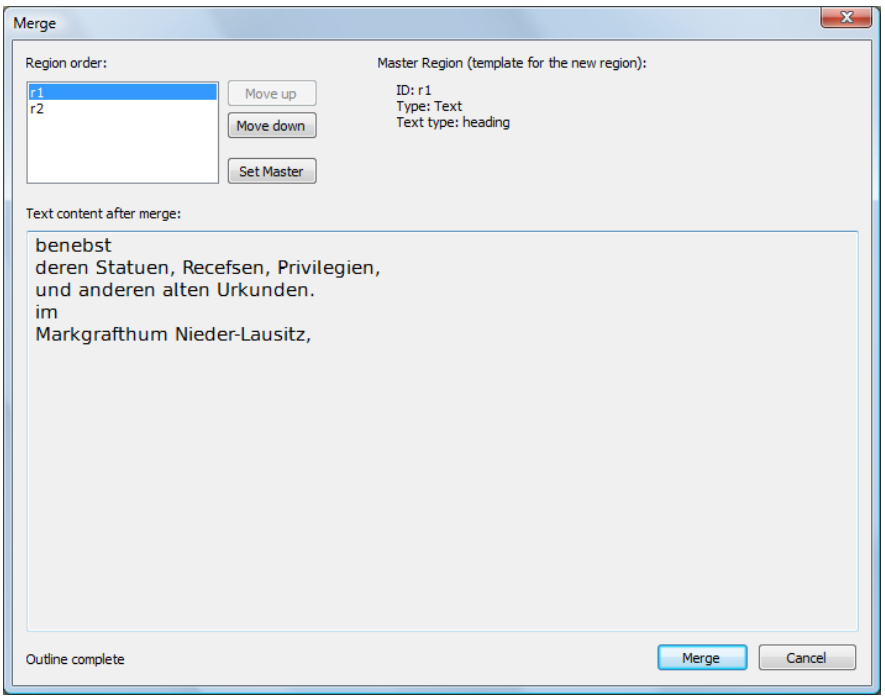


Merging Regions

To merge several regions into one region:

- Select the regions (see section ‘Selecting Regions’)
- Click ‘Merge’ in the toolbar panel called ‘Correction’ to open the Merge dialog (keyboard shortcut M)





- Optional: Select which region to use as template for the merged region
  - Select the region in the list box on the top left (properties are displayed on the right)
  - Click ‘Set Master’
- Optional: Change the region order that is used to merge the text content
  - Select a region in the list box on the top left
  - Click ‘Move up’ or ‘Move down’
  - Check the text preview at the bottom
- Click ‘Merge’ to finish (the merge button becomes active as soon as the outline of the new region has been calculated)

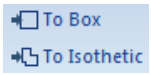
Existing text lines are reassigned to the new region.

Example:

Simplifying Region Outlines

To convert an outline of one or multiple regions:

- Select the region(s) (see section ‘Selecting Regions’)
- Click ‘To Box’ or ‘To Isothetic’ in the toolbar panel called ‘Correction’ (keyboard shortcuts Ctrl+B and Ctrl+I)



Example (left: original outline; middle: converted to box; right: converted to isothetic polygon):

Creating Text Regions from Text Lines (Bottom-up)

It is possible to create a new text region for selected text lines. This tool can be used to correct regions that have too many or too few text lines (over/under segmentation).

This tool is explained in chapter ‘Marking Text Lines’

Automatic Page Analysis and Text Recognition (OCR)

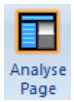
The open source OCR engine Tesseract has been integrated into Aletheia for automatic page analysis and text recognition. Other engines can be set up and used. For more information on Tesseract see:

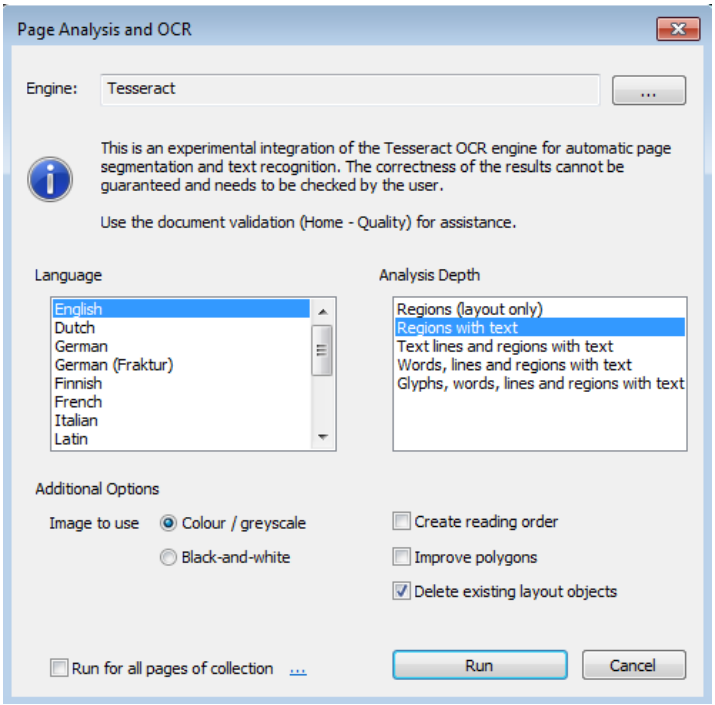
http://code.google.com/p/tesseract-ocr/
https://github.com/tesseract-ocr

Full Page Analysis

To analyse layout and text for the whole document page:

- Click ‘Analyse Page’ from the toolbar panel called ‘Auto’, a dialog opens.





- Select a language (select multiple languages using the CTRL key)
- Select the depth of the analysis:
  - Regions (layout only) – Just segments the document into layout regions. No text recognition is carried out.
  - Regions with text – Segments the document into regions and fills text regions with recognised text.
  - Text lines and regions with text – Segments the document into regions and fills text regions with recognised text. Text regions are further segmented into text lines.
  - Words, text lines and regions with text – Segments the document into regions and fills text regions with recognised text. Text regions are further segmented into text lines and lines

are segmented into words.
    - Glyphs, words, text lines and regions with text – Segments the document into regions and fills text regions with recognised text. Text regions are further segmented into text lines, lines are segmented into words and words are segmented into glyphs (characters).
- Choose the image to be used
- Decide if to keep existing layout objects (regions, lines, words and glyphs) and tick or untick the checkbox at the bottom accordingly.
- Click on 'Run'

Reading Order

Select this option to create a sequential reading order as calculated by Tesseract.

Improve Polygons

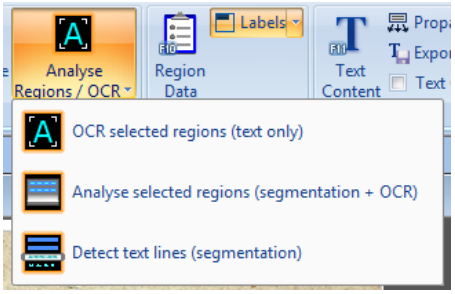
Select this option to recalculate the layout object polygons based on their child objects. This is most effective if words and/or glyphs are included in the page analysis.

Run for all Pages
Use this option to run the tool for all pages. See the chapter on Page Collections.

Note:
Aletheia 4 switched to Tesseract 4 as default OCR engine. Tesseract 3 is still available via the main settings or the "…" button in the OCR dialog.

Segmentation, text recognition and deep analysis for selected regions

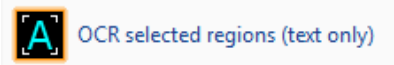
To process region(s):
- Select the region(s)
- Open the "Analyse regions / OCR" menu
- Choose one of the options (explained in the following subsections)

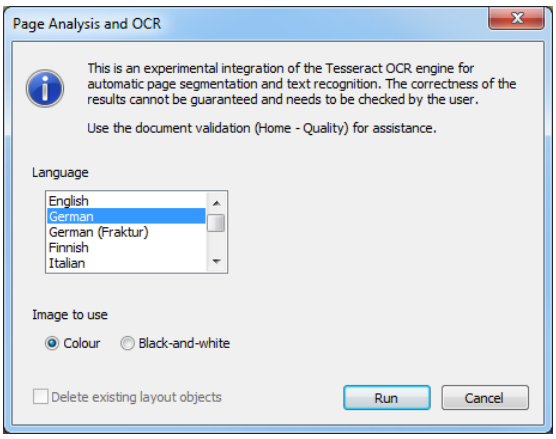


Recognising the Text for a Region (OCR)

To recognise the text content of a selected region:

- Select the region(s)
- Click 'OCR selected regions' from the toolbar panel called 'Auto' (keyboard shortcut 'o'), a dialog opens.





- Select a language
- Choose the image to be used
- Click on ‘Run’

Layout analysis and text recognition (OCR) for regions

To run a deep analysis of selected regions using an OCR engine:

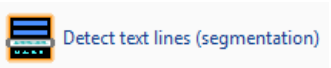
- Select the region(s)
- Click on ‘Analyse selected regions’; a dialog opens

- Select language, analysis depth and image to use
- Click on ‘Run’

Text line detection

To segment selected regions into text lines (polygonal outlines):

- Select the region(s)
- Click on ‘Detect text lines’

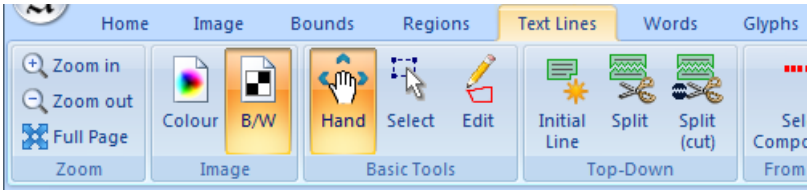


- (switch to Text Lines to see the results)

Marking Text Lines

To create, view and edit text lines:

- Activate the category 'Text Lines' within the toolbar (keyboard shortcut F7)

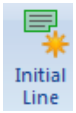


Note:
Logically, text lines are sub objects of text regions. However, text lines can also exist without a parent text region (pending text lines). Nevertheless, a final document layout should not contain such pending text lines (the document validator reports these as errors).

Defining Text Lines by Splitting (Top-Down)

To create text lines by starting from a text region (top-down approach):

- Activate 'Initial Line' in the toolbar panel called 'Top-Down' (keyboard shortcut 1) (press SHIFT and CTRL to create initial lines for all text regions at once)



- Click on a text region (a line spanning the whole region is created)

Following steps are only necessary if the region consists of more than one line.

- Activate 'Split' or 'Split (cut)' (keyboard shortcuts 2 and 3) (see the explanation below for what is the difference between the tools)

- Mark a split line using one of the following options:
  - If the lines are straight and there is enough space between two lines:
    - Position the mouse cursor in the space between two text lines
    - A horizontal line indicates where the initial line will be split

    - Click to execute the split

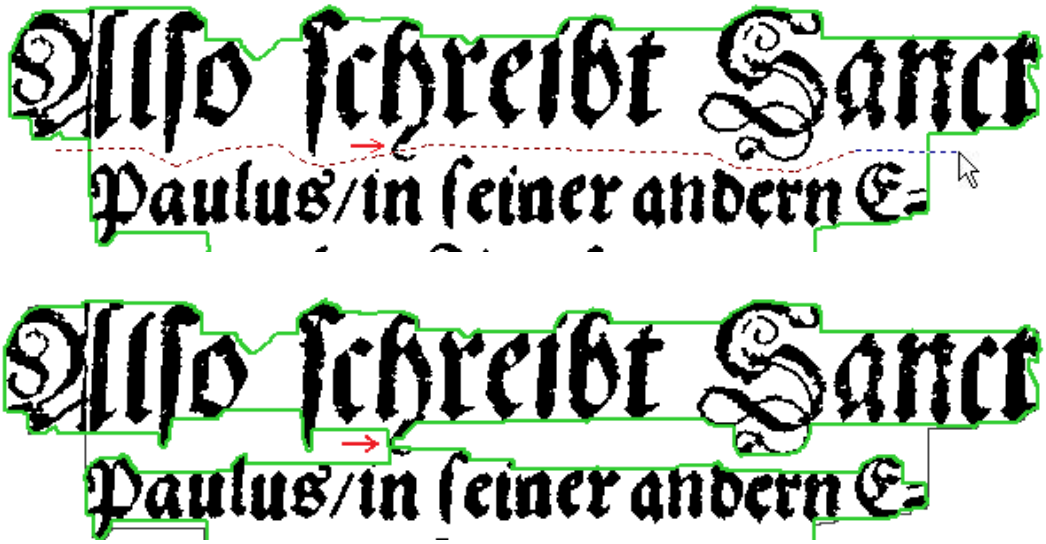
    Note: Press and hold CTRL or SHIFT to switch to a vertical split line

  - Alternatively draw a split line:
    - Start outside the region and repeatedly click left to define the line, finish with right or double click (latter adds a final point):

Preserve Components vs. Cut Components

The difference between the two splitting tools only becomes apparent if the split line touches a connected component (black object) either by accident or because the lines are connected.

The result for the ‘Cut’ tool looks like:



The component the split line crossed is ‘cut’ into two pieces. The upper part belongs to the first text line and the lower part to the second text line. Although in this example that behaviour is not wanted, there are examples where cutting is necessary:

The result using the first tool (that preserves components) looks like this:

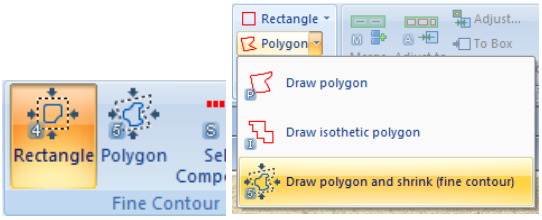
The component is not cut into two pieces.

Defining Text Lines using Contour Detection

Single text lines can be created by semi-automated contour detection.

To create a text line with contour detection:

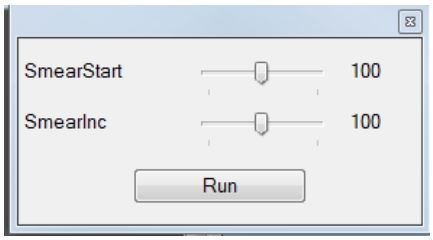
- Activate Rectangle or Polygon Fine Contour tool (keyboard shortcuts 4 and 5)



- Roughly specify the outline of the text line using the selected drawing tool (see the sections “Drawing Polygons” and “Drawing Rectangles”).

The recalculate the contour of existing text lines:

- Select the text lines of interest
- Activate any Fine Contour tool (keyboard shortcuts 4 and 5)
- Click on Run within the tool dialog

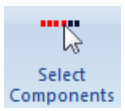


Defining Text Lines by Selecting Connected Components

Single text lines can also be defined by selecting all connected components (black objects) belonging to the line.

To create a text line with contour detection around a selection of components:

- Activate “Select Components” (keyboard shortcut S)



- Select components by
  - Left click on a component OR
  - Dragging a rectangle around components
- (Selected components are marked red)
- Add other components by clicking left or dragging a rectangle with the 'Ctrl' key pressed. Use ‘Shift’ and the mouse to toggle the current selection.
- Click Create Line (keyboard shortcut C)

Example:

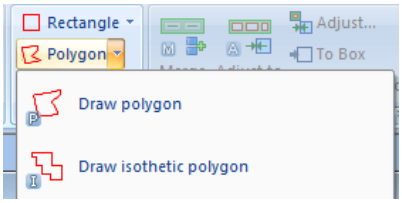

Joh. Christian Edler von Schmidt


Joh. Christian Edler von Schmidt

Drawing Text Lines Manually

To create a text line by drawing its outline:

- Activate Rectangle, Polygon or Isothetic from the toolbar panel called “Draw Contour” (keyboard shortcuts R, P and I)



- Draw the intended outline (see sections “Drawing Polygons” and “Drawing Rectangles”)

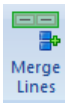
Manually marked text lines are automatically assigned to a parent region using their location within the document.

Correcting Text Lines

Merging Text Lines

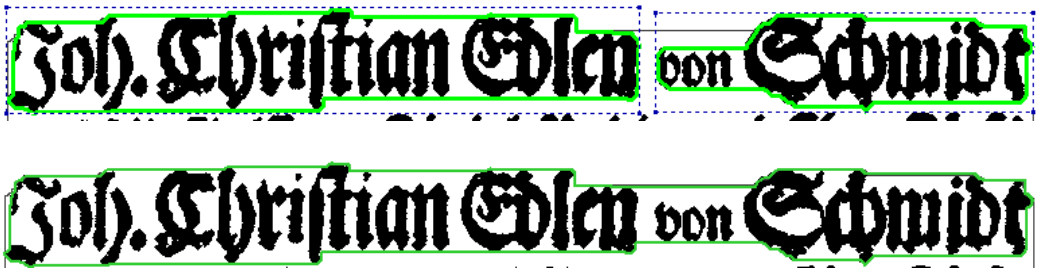
To merge wrongly split text lines:

- Select the text lines (line parts) that should be merged (see section ‘Selecting Regions’)
- Click ‘Merge Lines’ in the toolbar panel called ‘Correction’ (keyboard shortcut M)



- (Check the text content of the new text line)

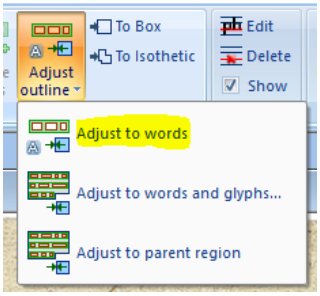
Example:



Adjusting Text Line Outlines to Child Text Objects

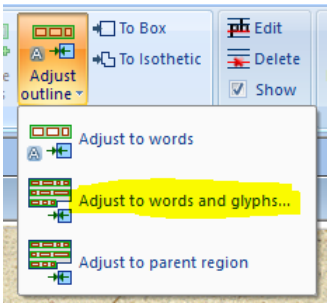
To adjust the outline of a text line to its word objects:

- Select the text line (see section ‘Selecting Regions’)
- Click ‘Adjust to words’ in the toolbar panel called ‘Correction’ (keyboard shortcut A)



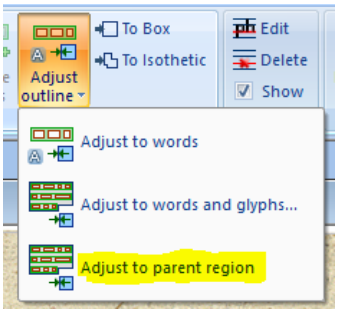
To adjust the outline of a text line to its child glyphs and words:

- Select the text line (see section ‘Selecting Regions’)
- Click ‘Adjust to words and glyphs’ in the toolbar panel called ‘Correction’



To limit the outline of a text line to the outline of its parent text region:

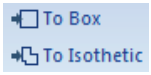
- Select the text line (see section ‘Selecting Regions’)
- Click ‘Adjust to parent region’ in the toolbar panel called ‘Correction’



Simplifying Text Line Outlines

To convert an outline of one or multiple text lines:

- Select the text line(s) (see section ‘Selecting Regions’)
- Click ‘To Box’ or ‘To Isothetic’ in the toolbar panel called ‘Correction’ (keyboard shortcuts Ctrl+B and Ctrl+I)

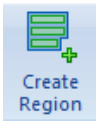


Creating Text Regions from Text Lines (Bottom-up)

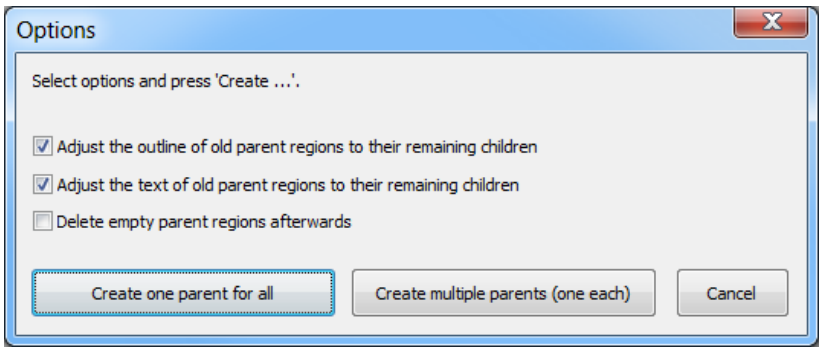
It is possible to create a single new text region for selected text lines or create multiple new text regions – one for each text line. This tool can be used to correct regions that have too many or too few text lines (over/under segmentation).

To create a single parent text region for selected text lines:

- Select the text line(s) (see section 'Selecting Regions')
- Click 'Create Region' in the toolbar panel called 'Bottom-up' (keyboard shortcut B)



- Optional: Choose options:
  - Tick the first checkbox if the selected text lines are already assigned to a text region and you want the outline of this (old) region to be adjusted to its remaining child text lines.
  - Tick the second checkbox if the selected text lines are already assigned to a text region and you want the text content of this (old) region to be adjusted to the text of its remaining child text lines.
  - Tick the third checkbox if the selected text lines are already assigned to a text region and you want this region to be deleted in case it has no more child text lines after the operation.



- Click 'Create one parent for all' to create the region

Example:

There are two text regions with overall 4 text lines. A check revealed that there should be only one region

for all lines. One possible solution would be to delete the two regions and create a new one. However, then the text lines with all their information would be lost.
Instead, the ‘Create Text Region’ tool can be used. All four text lines have to be selected:

In the dialog that pops up, all three options are checked. The result looks like this:

There is a new text region containing the four text lines. The old regions are deleted.

Note:
As a new text region is created, it is initialized with the default properties. The properties of old regions are not transferred to the new region. Only the text is automatically inserted into the new region (if the lines contain text).

To create one parent text regions for **each** selected text line:

- Follow the same steps as above
- Click ‘Create multiple parents’

Creating Text Lines from Words (Bottom-up)

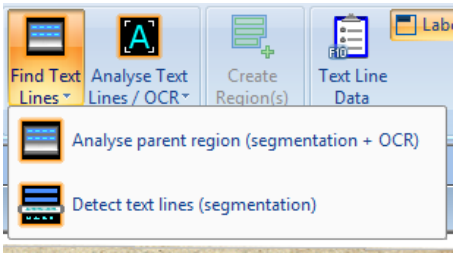
It is possible to create a new text line for selected words. This tool can be used to correct lines that have too many or too few words (over/under segmentation).

This tool is explained in chapter ‘Marking Words’.

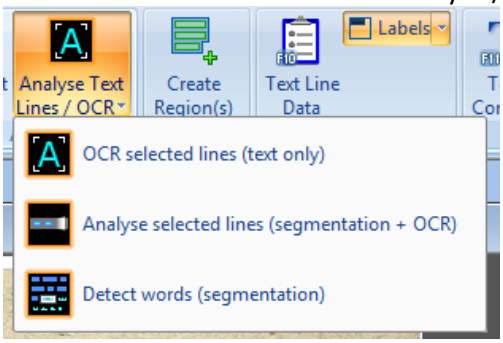
Analysis, Detection and Text Recognition

There are two main sets of automated tools:

1. Tools that create / detect text lines



2. Tools that work on text lines and analyse / recognise their content



Layout analysis and text recognition for a region via OCR engine

The open source OCR engine Tesseract has been integrated into Aletheia for automatic page analysis and text recognition. Other engines can be set up. For more information on Tesseract see:

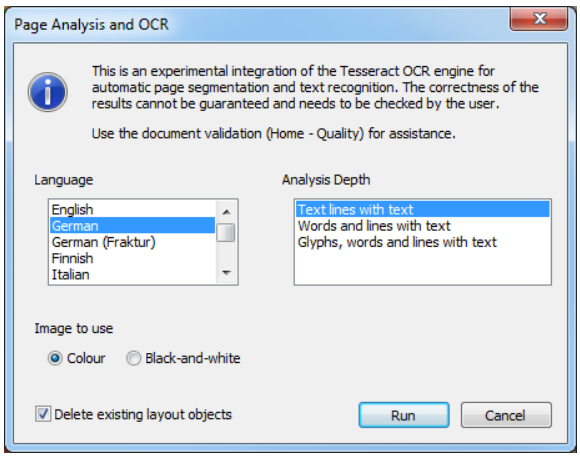
http://code.google.com/p/tesseract-ocr/
https://github.com/tesseract-ocr

Analysing a Region

To analyse layout (text lines, words, glyphs) and text for a selected region:

- Activate ‘Analyse parent region’ from the toolbar panel called ‘Auto’

- Click on an existing text region, a dialog opens.



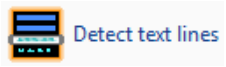
- Select a language
- Select the depth of the analysis:
    - Text lines with text – Segments the region into text lines and fills in the text content.
    - Words and lines with text – Segments the region into text lines and fills in the text content. Text lines are further segmented into words.
    - Glyphs, words and lines with text – Segments the region into text lines and fills in the text content. Text lines are further segmented into words and words are segmented into glyphs (characters).
- Choose the image to be used
- Decide if to keep existing layout objects (lines, words and glyphs) and tick or untick the checkbox at the bottom accordingly.
- Click on 'Run'

Text line detection

Note: This tool is experimental. The correctness of the results cannot be guaranteed and has to be verified manually.

To auto-detect text lines (polygonal outlines) for a region:

- Activate 'Detect Lines' in the toolbar panel called 'Auto' (keyboard shortcut D)



- Click inside a text region to start the detection.
- Optional: Check the results:
    - Select the first detected text line
    - Press Page Down repeatedly until the last line has been reached

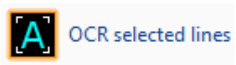
Note that the text content of the selected region is not automatically propagated to the detected text lines. However, Aletheia offers several tools dedicated to this task (see section "Compose and Extract Text" in chapter "Region Properties and Text Content").

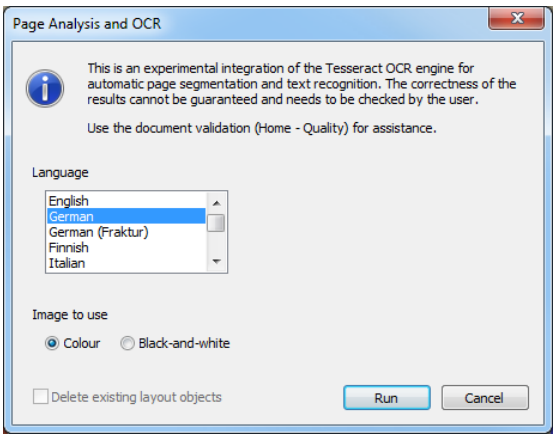
Example:

Recognising the text for text lines (OCR)

To recognise the text content of selected text line(s):

- Select the text line(s)
- Click 'OCR selected lines' from the toolbar panel called 'Auto' (keyboard shortcut 'o'), a dialog opens.



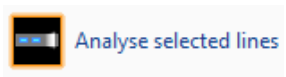


- Select a language
- Choose the image to be used
- Click on 'Run'

Analysing the content of a text line using an OCR engine

To analyse layout (words, glyphs) and text for selected text line(s):

- Select the text lines
- Click on 'Analyse selected lines' from the toolbar panel called 'Auto'; a dialog opens



- Select language, analysis depth and image to use
- Click on 'Run'
- (Check the results under Words)

Word detection for selected text lines

To auto-detect words (polygonal outlines) for text line(s):

- Select the text line(s)
- Click on 'Detect words in the toolbar panel called 'Auto'

- (Check the results under Words)

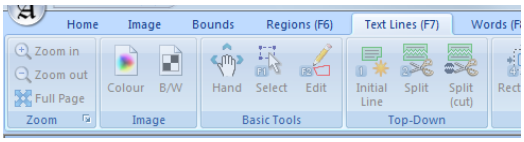
Marking Baselines

The baseline is a virtual line that connects the bottom most points (excluding descenders) of all characters of a text line:

(Source: Wikipedia)

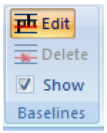
To mark or edit baselines:

- Switch to the “Text Lines” toolbar tab



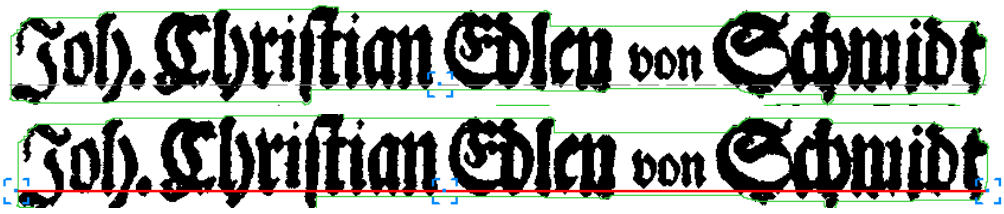
- (If not done already, mark the text lines for which you want to add the baseline)

- Activate ‘Edit’ from the ‘Baselines’ toolbar panel



- Click inside a text line to add baseline points
  - For assistance the cursor will snap to the bottom edge of a glyph (if a black-and-white image is available). To disable this feature, press CTRL wile clicking.

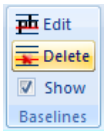
  - It is possible to add only one point (useful when the text line is straight). Aletheia will then add two additional points on saving or when another tool is selected.



- Click on an existing baseline point to delete it

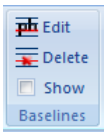
To delete baselines:

- Switch to the default tool (hand or select)
- Select the text lines from which you want to delete the baselines
- Click on ‘Delete’ in the ‘Baselines’ toolbar panel



To hide baselines:

- Untick ‘Show’ in the ‘Baselines’ toolbar panels to hide all baselines from view (Note: Baselines are always visible when the ‘Edit’ tool is active.)

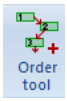


Text Line Order

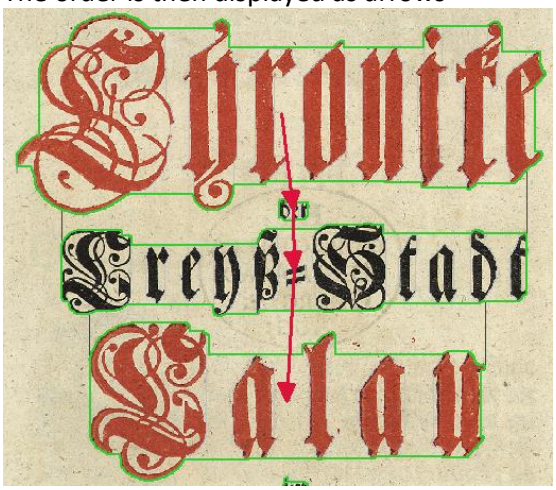
The order of text lines is usually defined by the text region the lines belong to. A text region has attributes for orientation (rotation) and text line order (top-to-bottom, bottom-to-top, …).

To manually specify the order of text lines:

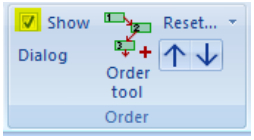
- Switch to the Text Line tab
- Activate the Interactive Tool



- Option 1:
  - Click on each text line of one region in the desired order
- Option 2:
  - Click and drag to “draw” a path with the mouse, connecting all text lines in the desired order
- The order is then displayed as arrows

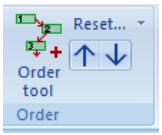


- Show or hide the order using the “Show” checkbox or by opening the text line order dialog



Editing the text line order

Using the toolbar controls



To add text lines to the sequential order:

- Activate the Order Tool
- Click on each text line
- (Optional: Hold the CTRL key while clicking to start a new sequence)

To remove text lines from the order:

- Select the text line(s)
- Click on “Remove” in the Reset… drop-down menu

OR

- Activate the Interactive Tool
- Right-click on text lines to remove

To move a text line forward or backward in the order:

- Select the text line
- Click on the arrows in the Order toolbar panel

To delete one text line sequence:

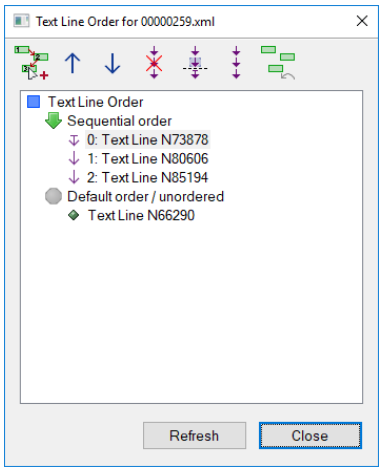
- Select a text line of the sequence
- Click on “Reset the selected sequence” in the Reset… drop-down menu

To delete all text line ordering of the current page:

- Click on “Reset all” in the Reset… drop-down menu
- Confirm

Using the text line order dialog

The dialog gives more control for editing complex text line order. Select one text line to update the dialog. All sibling text lines are listed in the tree.



To add a text line to the sequential order:

- Drag a text line time from “Default order”
    - Drop it on “Sequential order” to add it to the end of the sequence
      OR
    - Drop it on an item under Sequential Order to insert it at that position

To remove a text line from the sequential order:

- Drag the item from “Sequential order” to “Default order”
  OR
- Select the text line and click on the remove button 

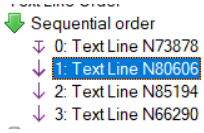
To remove all text lines from the sequential order:

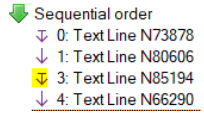
- Use the “Reset” button

To move and item up or down in the sequential order:

- Select the item and use the arrow buttons
  OR
- Drag the item to another item in the sequential order (the other item will be shifted down)

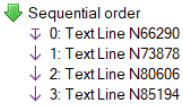
To separate a sequence into two sequences:

- Select text line
  

- Click on the separate button
  
- The sequence is separated after the selected text line
  

To join sequences:

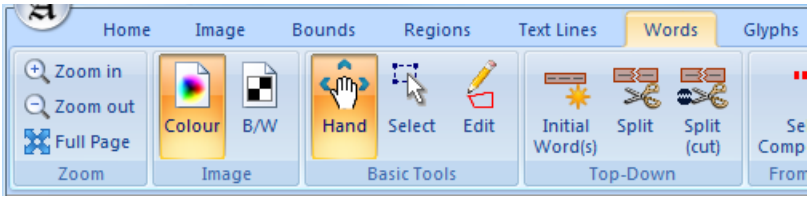
- Click on the join button



Marking Words

To create, view and edit words:

- Activate the category 'Words' within the toolbar (keyboard shortcut F8)



Note:
Logically, words are sub objects of text lines. However, words can also exist without a parent text line (pending words). Nevertheless, a final document layout should not contain such pending objects (the document validator reports these as errors).

Defining Words by Splitting (Top-Down)

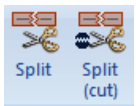
To create words by starting from a text line (top-down approach):

- Activate 'Initial Words' in the toolbar panel called 'Top-Down' (keyboard shortcut 1) (press SHIFT and CTRL to create initial words for all text lines at once)

- Click on a text region (words spanning one whole text line each are created)

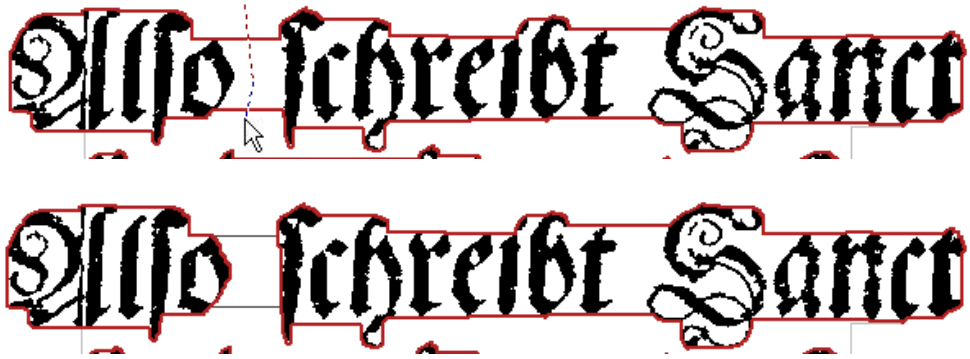
Following steps are only necessary if a text line consists of more than one word.

- Activate 'Split' or 'Split (cut)' (keyboard shortcuts 2 and 3) (see the explanation below for what is the difference between the tools)



- Mark a split line using one of the following options:
  - If the words are not connected and there is enough space between two words:
    - Position the mouse cursor in the space between two words
    - A vertical line indicates where the initial word will be split
    - Click to execute the split
  - Alternatively draw a split line:
    - Start outside the word to split and repeatedly click left to define the line, finish with right or double click (latter adds a final point) (see chapter 'Drawing Polgons and Polylines')

Example:



Preserve Components vs. Cut Components

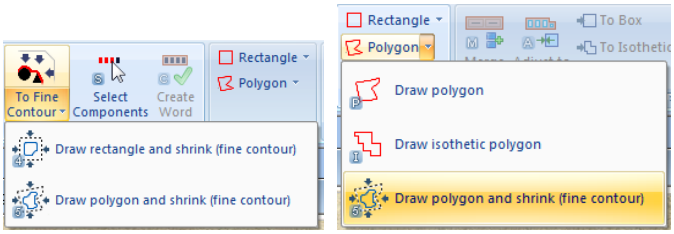
The ‘Cut’ splitting tool cuts connected components into two pieces if the splitting line crosses them, either by accident or because the glyphs of two words are connected. The first tool (that preserves components) doesn’t cut the connected components. See the chapter ‘Defining Text Lines’ for an example.

Defining Words using Contour Detection

Single words can be created by semi-automated contour detection.

To create a word with contour detection:

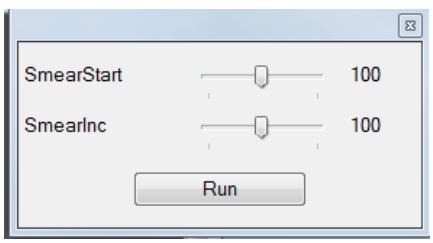
- Activate Rectangle or Polygon Fine Contour tool (keyboard shortcuts 4 and 5)



- Roughly specify the outline of the word using the selected drawing tool (see the sections “Drawing Polygons” and “Drawing Rectangles”).

The recalculate the contour of existing words:

- Select the words of interest
- Activate any Fine Contour tool (keyboard shortcuts 4 and 5)
- Click on ‘Run’ within the tool dialog

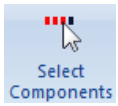


Defining Words by Selecting Connected Components

Single words can also be defined by selecting all connected components (black objects) belonging to the word.

To create a word with contour detection around a selection of components:

- Activate “Select Components” (keyboard shortcut S)



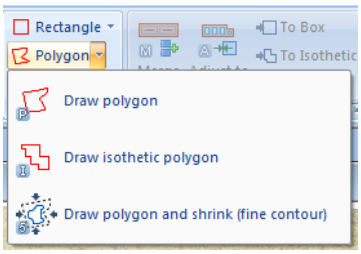
- Select components by
  - Left click on a component OR
  - Dragging a rectangle around components
- (Selected components are marked red)
- Add other components by clicking left or dragging a rectangle with the 'Ctrl' key pressed. Use ‘Shift’ and the mouse to toggle the current selection.
- Click 'Create Word' (keyboard shortcut C)

Example:

Drawing Words Manually

To create a word by drawing its outline:

- Activate Rectangle, Polygon or Isothetic polygon from the toolbar panel called “Draw Contour” (keyboard shortcuts R, P and I)



- Draw the intended outline (see sections 'Drawing Polygons' and 'Drawing Rectangles')

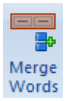
Manually marked words are automatically assigned to a parent text line using their location within the document.

Correcting Words

Merging Words

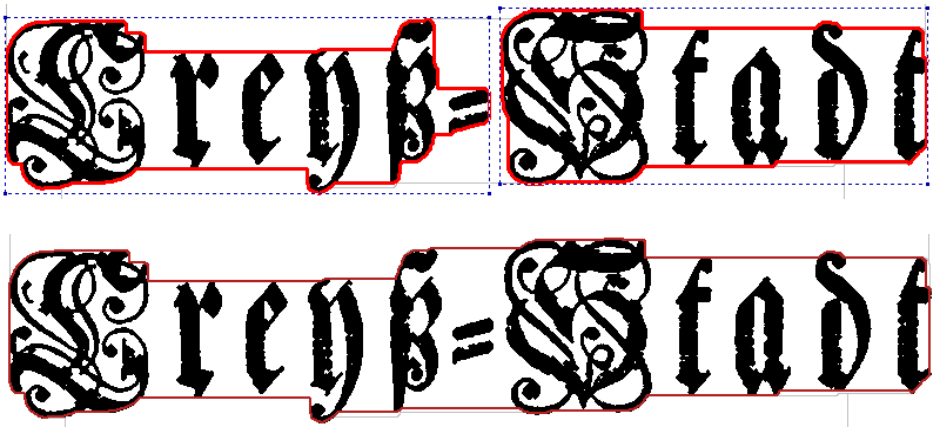
To merge wrongly split words:

- Select the words (parts of a word) that should be merged (see section “Selecting Regions”)
- Click “Merge Words” in the toolbar panel called Correction (keyboard shortcut M)



- (Check the text content of the new word)

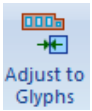
Example:



Adjusting Word Outlines to Glyphs

To adjust the outline of a word to its glyph objects:

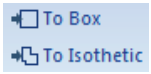
- Select the word (see section “Selecting Regions”)
- Click “Adjust to glyphs” in the toolbar panel called Correction (keyboard shortcut A)



Simplifying Word Outlines

To convert an outline of one or multiple words:

- Select the word(s) (see section ‘Selecting Regions’)
- Click ‘To Box’ or ‘To Isothetic’ in the toolbar panel called ‘Correction’ (keyboard shortcuts Ctrl+B and Ctrl+I)



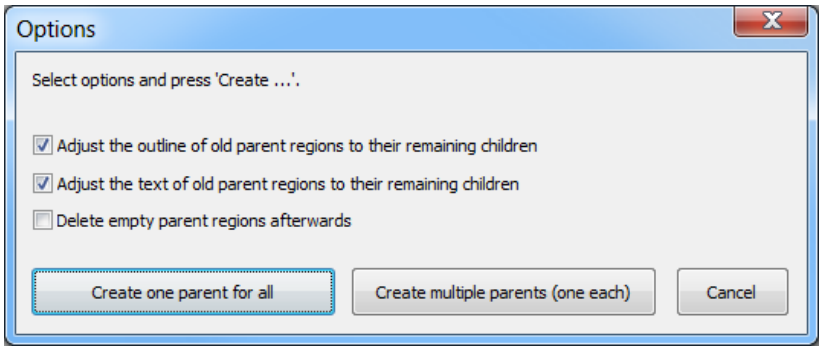
Creating Text Lines from Words (Bottom-up)

It is possible to create a single new text line for selected words or multiple new parent text lines – one for each word. This tool can be used to correct lines that have too many or too few words (over/under segmentation).

To create a single parent text line for selected words:

- Select the word(s) (see section “Selecting Regions”)
- Click “Create Text Line” in the toolbar panel called Bottom-up (keyboard shortcut B)

- Optional: Choose options:
  - Tick the first checkbox if the selected words are already assigned to a text line and you want the outline of this (old) text line to be adjusted to its remaining child words.
  - Tick the second checkbox if the selected words are already assigned to a text line and you want the text content of this (old) line to be adjusted to the text of its remaining child words.
  - Tick the third checkbox if the selected words are already assigned to a text line and you want this line to be deleted in case it has no more child words after the operation.



- Click ‘Create one parent for all’ to create the text line

To create one parent text line for **each** selected word:

- Follow the same steps as above
- Click ‘Create multiple parents’

Creating Words from Glyphs (Bottom-up)

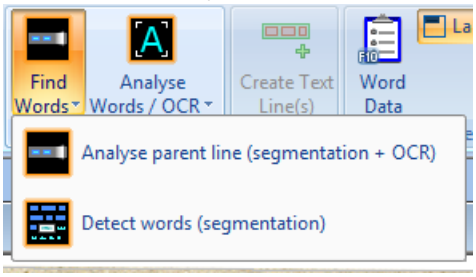
It is possible to create a new word for selected glyphs. This tool can be used to correct words that have too many or too few glyphs (over/under segmentation).

This tool is explained in chapter ‘Marking Glyphs’.

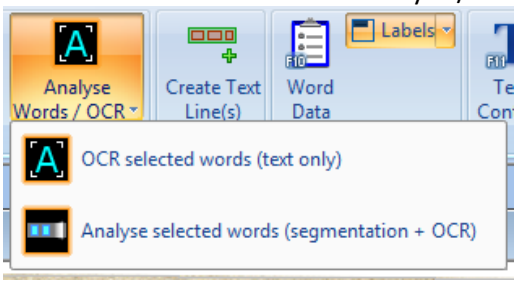
Analysis, Detection and Text Recognition

There are two main sets of automated tools:

1. Tools that create / detect words



2. Tools that work on words and analyse / recognise their content



Layout analysis and text recognition for a parent text line using an OCR engine

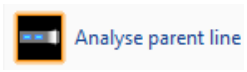
The open source OCR engine Tesseract has been integrated into Aletheia for automatic page analysis and text recognition. Other engines can be set up. For more information on Tesseract see:

http://code.google.com/p/tesseract-ocr/
https://github.com/tesseract-ocr

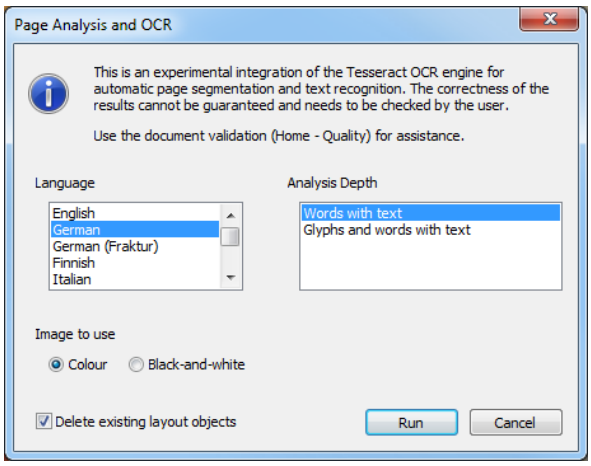
Analysing a Text Line

To analyse layout and text for a selected text line:

- Activate ‘Analyse Line’ from the toolbar panel called ‘Auto’



- Click on an existing text line, a dialog opens.



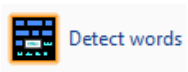
- Select a language
- Select the depth of the analysis:
  - Words with text – Segments the text line into words and fills in the text content.
  - Glyphs and words with text – Segments the text line into words and fills in the text content. Words are further segmented into glyphs (characters).
- Choose the image to be used
- Decide if to keep existing layout objects (words and glyphs) and tick or untick the checkbox at the bottom accordingly.
- Click on 'Run'

Word Detection

Note: This tool is experimental. The correctness of the results cannot be guaranteed and has to be verified manually.

To auto-detect words for a region with text lines:

- Activate 'Detect Words' in the toolbar panel called 'Auto' (keyboard shortcut D)



- Click inside a text region to start the detection.
- Optional: Check the results:
  - Select the first detected word
  - Press Page Down repeatedly until the last word has been reached

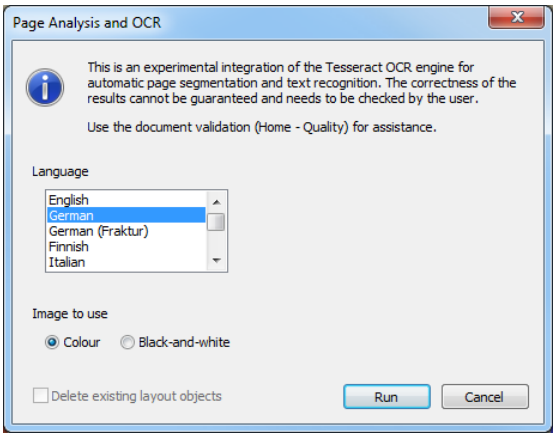
Note that the text content of the region or text lines is not automatically propagated to the detected words. However, Aletheia offers several tools dedicated to this task (see section "Compose and Extract Text" in chapter "Region Properties and Text Content").

Example:

Recognising the text for selected word(s) (OCR)

To recognise the text content of selected word(s):

- Select the word(s)
- Click ‘OCR Word’ from the toolbar panel called ‘Auto’ (keyboard shortcut ‘o’), a dialog opens.



- Select a language
- Choose the image to be used
- Click on ‘Run’

Layout analysis and text recognition for selected words using an OCR engine

To analyse layout (glyphs) and text for selected words:

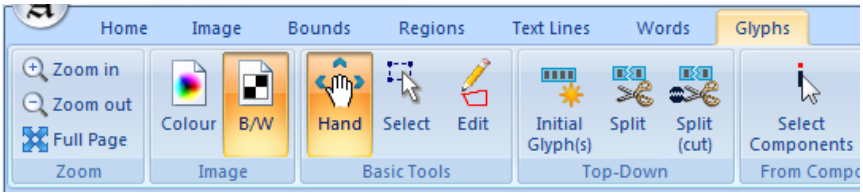
- Select the words
- Click ‘Analyse selected words’ from the toolbar panel called ‘Auto’

- (Check the results under Glyphs)

Marking Glyphs

To create, view and edit glyphs (symbols / characters):

- Activate the category Glyphs within the toolbar (keyboard shortcut F9)



Note:
Logically, glyphs are sub-objects of words. However, glyphs can also exist without a parent word (pending glyphs). Nevertheless, a final document layout should not contain such pending objects (the document validator reports these as errors).

Defining Glyphs by Splitting (Top-Down)

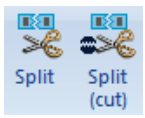
To create glyphs by starting from a word (top-down approach):

- Activate “Initial Glyphs” in the toolbar panel called Top-Down (keyboard shortcut 1) (press SHIFT and CTRL to create initial glyphs for all words at once)

- Click on a text line (glyphs spanning one whole word each are created)

Following steps are only necessary if a word consists of more than one glyph.

- Activate ‘Split’ or ‘Split (cut)’ (keyboard shortcuts 2 and 3) (see the explanation below for what is the difference between the tools)



- Mark a split line using one of the following options:
  - If the glyphs are not connected and there is enough space between two glyphs:
    - Position the mouse cursor in the space between two glyphs
    - A vertical line indicates where the initial glyph will be split
    - Click to execute the split

  - Alternatively draw a split line:
    - Start outside the glyph to split and repeatedly click left to define the line, finish with right or

double click (latter adds a final point) (see chapter “Drawing Polygons and Polylines”)

Example:

Preserve Components vs. Cut Components

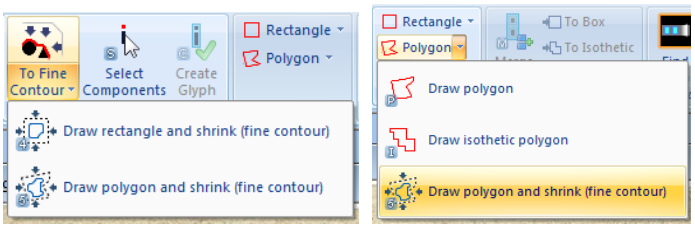
The ‘Cut’ splitting tool cuts connected components into two pieces if the splitting line crosses them, either by accident or because the glyphs are connected. The first tool (that preserves components) doesn’t cut the connected components. See the chapter ‘Defining Text Lines’ for an example.

Defining Glyphs using Contour Detection

Single glyphs can be created by semi-automated contour detection.

To create a glyph with contour detection:

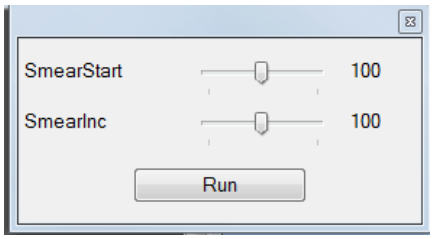
- Activate Rectangle or Polygon Fine Contour tool (keyboard shortcuts 4 and 5)



- Roughly specify the outline of the glyph using the selected drawing tool (see the sections “Drawing Polygons” and “Drawing Rectangles”).

The recalculate the contour of existing glyphs:

- Select the glyphs of interest
- Activate any Fine Contour tool (keyboard shortcuts 4 and 5)
- Click on Run within the tool dialog



Defining Glyphs by Selecting Connected Components

Single glyphs can also be defined by selecting all connected components (black objects) belonging to the glyph.

To create a glyph with contour detection around a selection of components:

- Activate 'Select Components' (keyboard shortcut S)

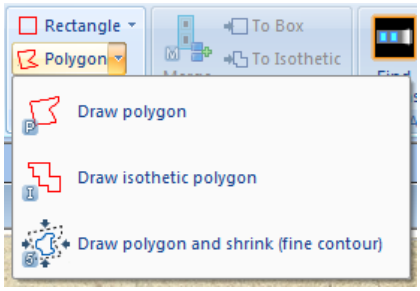
- Select components by
  - Left click on a component OR
  - Dragging a rectangle around components
- (Selected components are marked red)
- Add other components by clicking left or dragging a rectangle with the 'Ctrl' key pressed. Use ‘Shift’ and the mouse to toggle the current selection.
- Click 'Create Glyph' (keyboard shortcut C)

Example:

Drawing Glyphs Manually

To create a glyph by drawing its outline:

- Activate Rectangle, Polygon or Isothetic polygon from the toolbar panel called “Draw Contour” (keyboard shortcuts R, P and I)



- Draw the intended outline (see sections “Drawing Polygons” and “Drawing Rectangles”)

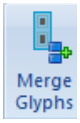
Manually marked glyphs are automatically assigned to a parent word using their location within the document.

Correcting Glyphs

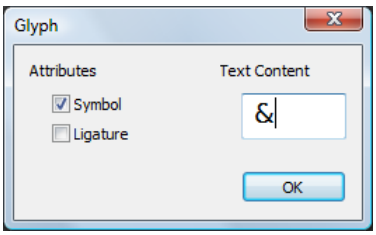
Merging Glyphs

To merge wrongly split glyphs:

- Select the glyphs (parts of a glyph) that should be merged (see section 'Selecting Regions')
- Click "Merge Glyphs" in the toolbar panel called Correction (keyboard shortcut M)



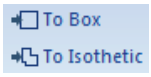
- (Check the text content of the new glyph)
- If the attributes of the merged glyphs conflict with each other, a dialog is shown to confirm the values:



Simplifying Glyph Outlines

To convert an outline of one or multiple glyphs:

- Select the glyph(s) (see section "Selecting Regions")
- Click "To Box" or "To Isothetic" in the toolbar panel called Correction (keyboard shortcuts Ctrl+B and Ctrl+I)

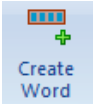


Creating Words from Glyphs (Bottom-up)

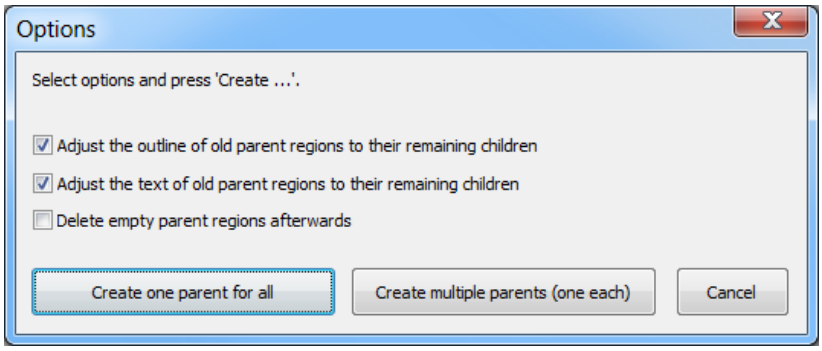
It is possible to create a single new word for selected glyphs or create multiple new words – one for each glyph. This tool can be used to correct words that have too many or too few glyphs (over/under segmentation).

To create a single parent word for selected glyphs:

- Select the glyphs(s) (see section 'Selecting Regions')
- Click 'Create Word' in the toolbar panel called 'Bottom-up' (keyboard shortcut B)



- Optional: Choose options:
  - Tick the first checkbox if the selected glyphs are already assigned to a word and you want the outline of this (old) word to be adjusted to its remaining child glyphs.
  - Tick the second checkbox if the selected glyphs are already assigned to a word and you want the text content of this (old) word to be adjusted to the text of its remaining child glyphs.
  - Tick the third checkbox if the selected glyphs are already assigned to a word and you want this word to be deleted in case it has no more child glyphs after the operation.



- Click ‘Create one parent for all’ to create the word

To create one parent word for **each** selected glyph:

- Follow the same steps as above
- Click ‘Create multiple parents’

Layout analysis and text recognition via OCR engine

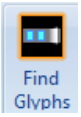
The open source OCR engine Tesseract has been integrated into Aletheia for automatic page analysis and text recognition. Other engines can be set up. For more information on Tesseract see:

http://code.google.com/p/tesseract-ocr/
https://github.com/tesseract-ocr

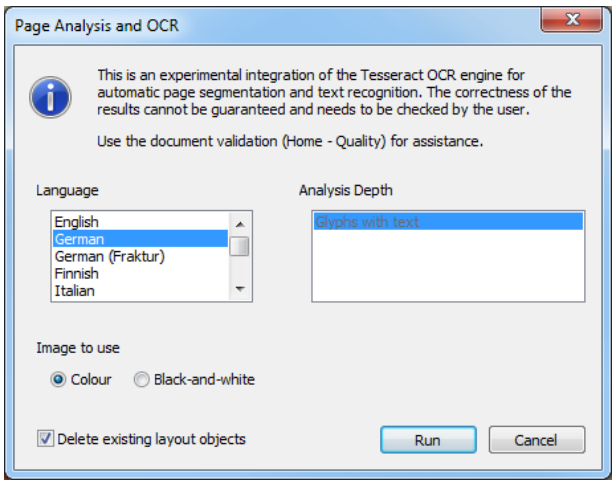
Analysing a Word

To analyse layout and text for a selected word:

- Activate ‘Find Glyphs’ from the toolbar panel called ‘Auto’



- Click on an existing word, a dialog opens.

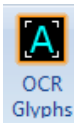


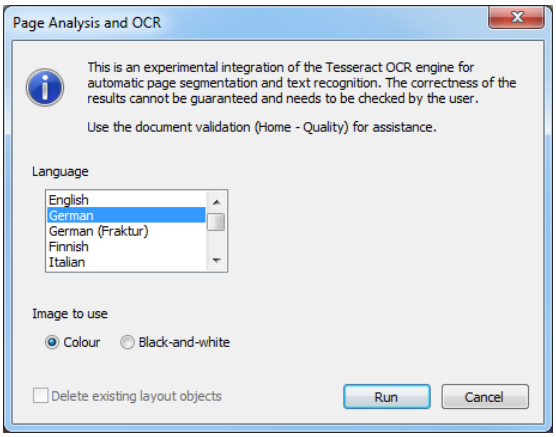
- Select a language
- Choose the image to be used
- Decide if to keep existing layout objects (glyphs) and tick or untick the checkbox at the bottom accordingly.
- Click on 'Run'

Recognising the text for selected glyphs (OCR)

To recognise the text content of selected glyph(s):

- Select the glyph(s)
- Click 'OCR Glyphs' from the toolbar panel called 'Auto' (keyboard shortcut 'o'), a dialog opens.





- Select a language
- Choose the image to be used
- Click on 'Run'

Drawing Tools

Note: A cross-hair aid at the current mouse position can be enabled at any time by pressing CTRL and SHIFT while moving the mouse cursor.

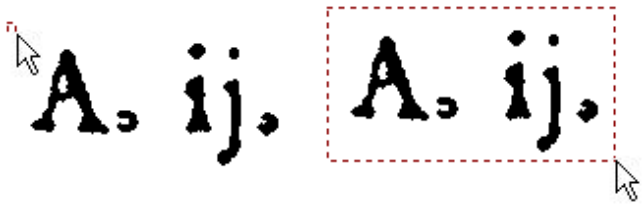
Drawing Rectangles

Rectangular regions can be used for the document image border, the print space and layout regions.

There are two ways to draw a rectangular region:

1. By dragging:
   - Move the mouse cursor to one corner of the desired rectangle
   - Press the left mouse button (and keep it pressed)
   - Drag the cursor to the opposite corner
   - Release the mouse button

2. By clicking twice
   - Move the mouse cursor to one corner of the desired rectangle
   - Click left
   - Move the cursor to the opposite corner
   - Click right (or click left again to redefine the first point)



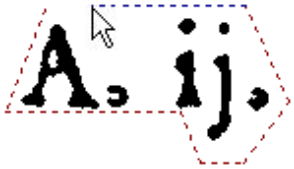
Drawing Polygons and Polylines

Polygonal regions can be used for the document image border, the print space and layout regions. Polylines are used for all splitting tools (splitting of text lines, words and glyphs).

To draw a polygon or polyline:

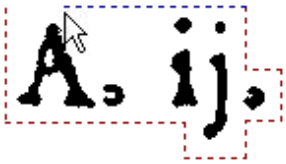
- Move the mouse cursor to the position of the first point of the desired polygon or polyline
- Click left
- Continue the first two steps for all succeeding points.
- Finish either by:
  - Double click left, to add another point and close the polygon or finish the polyline

  OR
  - Right click, to close the polygon / finish the polyline without adding another point



Isothetic Polygons

Isothetic polygons are a specialization of the general polygons. They only consist of vertical and horizontal lines.

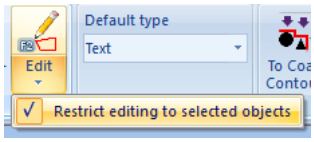


Editing Outlines

To edit an object outline (polygon):

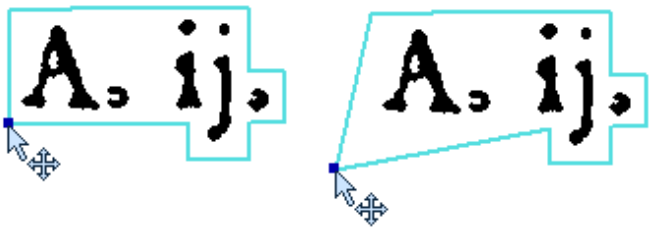
- Activate 'Edit' in the toolbar panel called ‘Basic Tools’ (keyboard shortcut F2)

Note:
To restrict the editing to one or multiple selected objects, select the respective option in the drop-down menu below the Edit toolbar button. If enabled, only the outlines of selected objects can be modified (see section Selecting Regions).



Moving Single Points

To move a polygon point, navigate the mouse cursor to the point until a blue marker appears over the point. Then press the left mouse button and drag the point to the desired position.



To ‘snap’ the point to the coordinates of the neighbour points, press Ctrl while dragging the point. That way lines can be made exactly horizontal or vertical.

Moving Multiple Points

To move multiple points:

- Select the points
    - Left click a single point or drag a selection frame around points
    - Press Ctrl to add points to the current selection
    - Press Shift to add to or remove points from the current selection
    - Double click a point to select all points of the polygon
    - Double click inside a region to select all points of the region
- Grab one point of the selection and drag the points to the intended position

Adding Points

To add a point to a polygon, navigate the mouse cursor to the desired position until a red plus appears above the polygon line. Then click left to add the point.

Deleting Points

To delete a polygon point, navigate the mouse cursor to the point until a blue marker appears over the point. Then press the delete key or use the 'Delete' button of the toolbar.

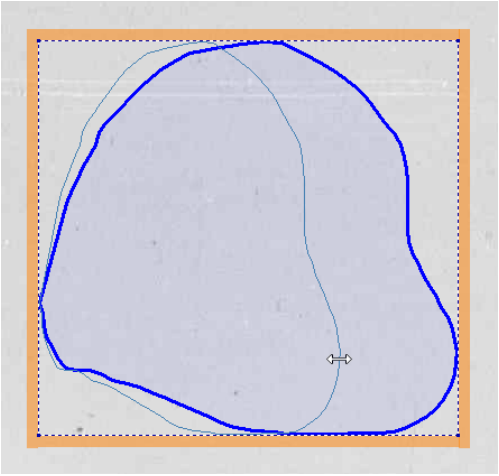
To delete multiple points at once, select the considered points and press delete or use the toolbar button.

Note: Deleting a selection of points has priority. If points are selected and the mouse hovers over another point, the delete action is carried out to the selection.

Resizing

A page object can be resized via the Move & Resize tool or via the Hand as well as the Select tool:

- Activate the Resize, Hand or Select tool
- Select a single object (e.g. a region)
- Place the mouse cursor on the selection rectangle (the cursor changes)
    - In case of the Hand or Select tool, press the SHIFT key at the same time
- Click and drag the mouse
- (Press CTRL while resizing to maintain the aspect ratio of the object)



The Resize tool can also be used to move a polygon:

- Select an object
- Activate the tool
- Position the mouse cursor inside (moving arrows appear)
- Click and drag with the mouse

Working with Regions and Other Page Objects

Selecting Regions

Switch to the 'Select' tool (keyboard shortcut F1) or to the ‘Hand’ tool (Space key).

Select a region by clicking left inside the region. Selected regions are marked by a dotted rectangle.

You can jump to the next or previous region by pressing the 'Page Down' or 'Page Up' key.

Multiple Selection

Press the Ctrl or Shift key while selecting a region to add it to the current selection.
To remove a region from the current selection press the Ctrl or Shift key and click on the region.

Selection Rectangle (‘Select’ tool only)

Click left and drag a rectangle around the regions to select multiple regions at once. Only regions that are

completely within the selection rectangle will be selected.
Use the Ctrl or Shift key to add/remove regions to/from the current selection.

Select all and select similar

To select all page objects of the current level (regions, text lines, words or glyphs):

- Press Ctrl+A or click on 'Select all' 

To select all page objects that are similar to the currently selected ones:

- Click on 'Select similar' 

Navigating to another object

To navigate to the next or previous object in a global order (reading order and/or top-to-bottom):

- (Select one object)
- Press Page Up or Page Down

To navigate to a neighbour object:

- (Select one object)
- Press CTRL + cursor key (left, right, up or down)

Deleting Page Objects

To delete an object, select it and press the 'Delete' key. A message box will show up. Select 'Yes' to confirm the delete operation. If an object contains sub-regions (children), an option is presented to keep these regions and delete the parent only. In this case, the sub-regions are temporarily assigned to an invisible parent object.

Reassigning Page Objects

It is possible to reassign existing objects to their correct parent regions, based on the position within the document. Following objects can be reassigned:

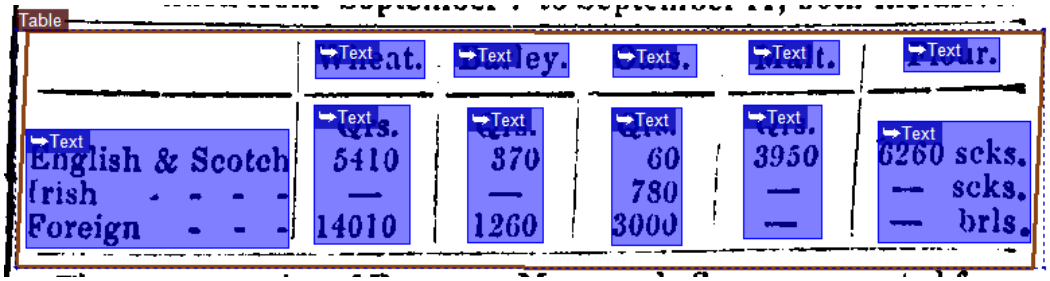
- A regions can be assigned to any other region (nested region, see next section)
- A text line can be assigned to a text region
- A word can be assigned to a text line
- A glyph can be assigned to a word

To reassign page objects:

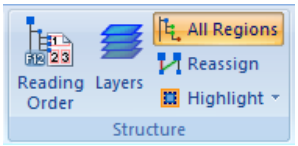
- Select the page objects to reassign (sub regions)
- Click 'Reassign' in the toolbar panel called 'Structure'

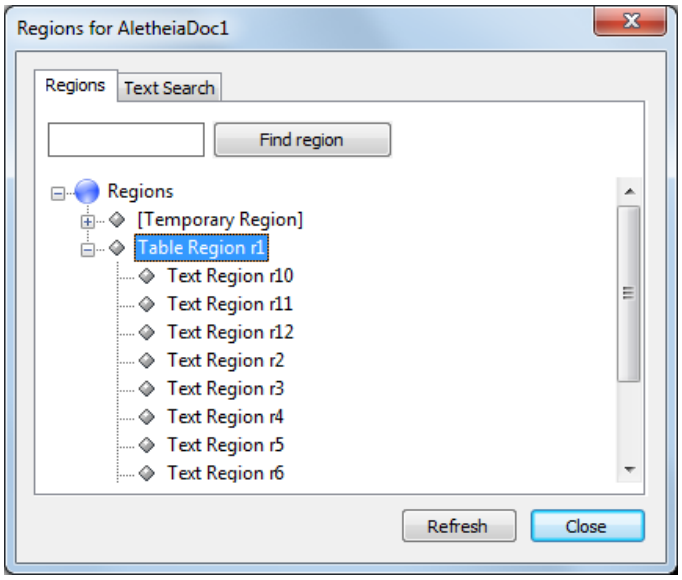
Nested Regions

Layout regions can be nested within other layout regions (parent-child relation). A table region, for example, can have text sub-regions to model the table cells:



Nested regions are indicated by the arrow at the beginning of the label. The relationship between the regions is also reflected in the region structure view (“All Regions”):





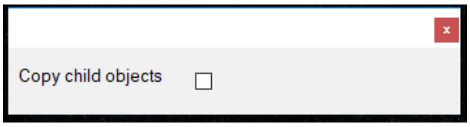
To un-nest a region:

- Open the “All Regions” dialog
- Drag the nested region to the “Regions” root element

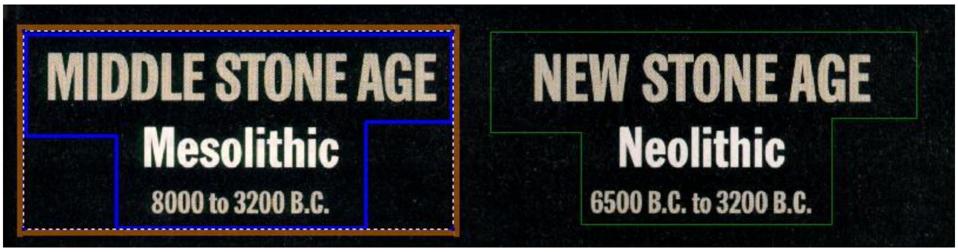
Copy & Paste

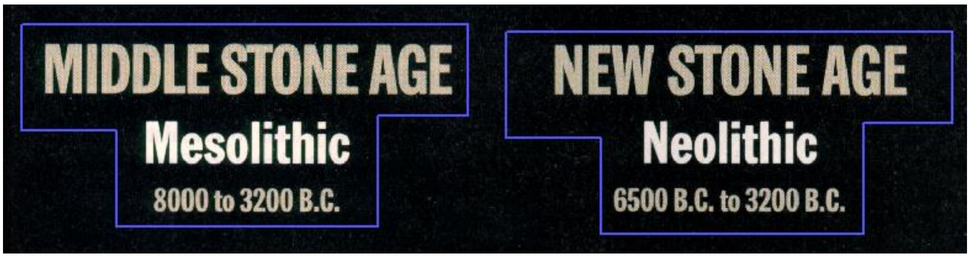
To copy and paste a region, text line, word or glyph:

- Select the object
- Press CTRL + C
- Press CTRL + V (activates paste tool)
- (Enable “Copy child objects” to include text lines if you paste a text region, for example)



- Position the outline where you want to paste
- Click to paste

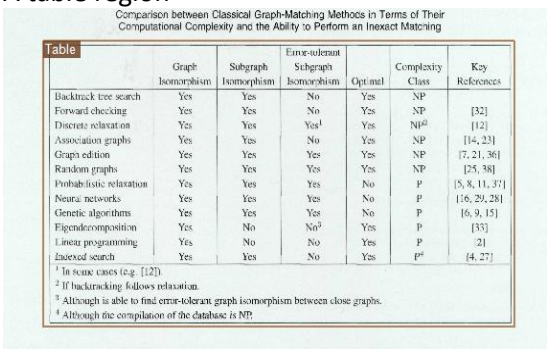




Tables

Tables are represented by:

- A table region



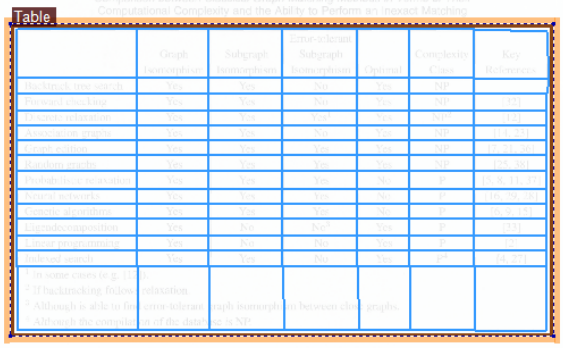
Comparison between Classical Graph-Matching Methods in Terms of Their Computational Complexity and the Ability to Perform an Inexact Matching

Table

	Graph Isomorphism	Subgraph Isomorphism	Error-tolerant Subgraph Isomorphism	Optimal	Complexity Class	Key References
Backtrack tree search	Yes	Yes	No	Yes	NP	
Forward checking	Yes	Yes	No	Yes	NP	[32]
Discrete relaxation	Yes	Yes	Yes[1]	Yes	NP[2]	[12]
Association graphs	Yes	Yes	No	Yes	NP	[14, 23]
Graph edition	Yes	Yes	Yes	Yes	NP	[7, 21, 36]
Random graphs	Yes	Yes	Yes	Yes	NP	[25, 38]
Probabilistic relaxation	Yes	Yes	Yes	No	P	[5, 8, 11, 37]
Neural networks	Yes	Yes	Yes	No	P	[16, 29, 28]
Genetic algorithms	Yes	Yes	Yes	No	P	[6, 9, 15]
Eigendecomposition	Yes	No	No[3]	Yes	P	[33]
Linear programming	Yes	No	No	Yes	P	[2]
Indexed search	Yes	Yes	No	Yes	P[4]	[4, 27]

[1] In some cases (e.g. [12]).
[2] If backtracking follows relaxation.
[3] Although is able to find error-tolerant graph isomorphism between close graphs.
[4] Although the compilation of the database is NP.

- A table grid (optional)



- Table cell regions (optional)

- Separator regions (optional)

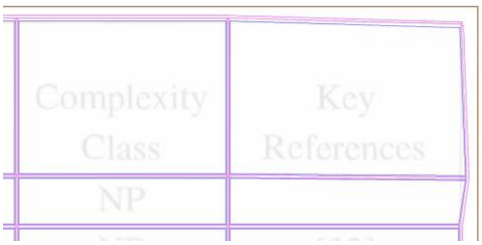


Table Regions

To create a table region:

- Switch to the Regions toolbar tab

- (Change the default type to Table)
- Create a region with any of the provided tools
- (Open the Region Attributes dialog and switch the type to Table)

To change table attributes:

- Select a table region
- Open the Region Attributes dialog

- Modify the attributes

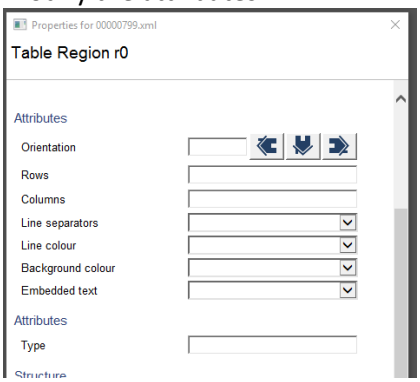
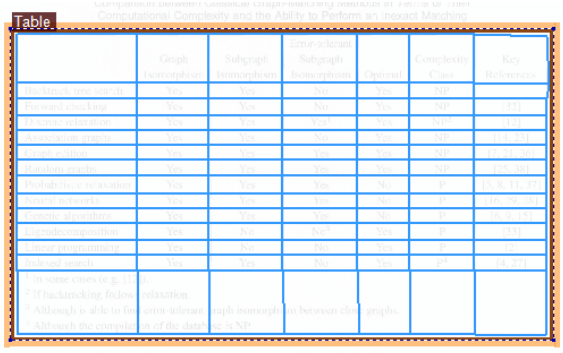


Table Grid

The grid defines the underlying row and column structure of a table. It does NOT define table cells. As such, it also does not include the concept of row or column spans (this is done when creating cells, see next section). The grid lines are guides for creating cell regions and separator regions.

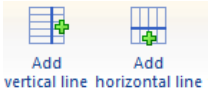


There are three ways to add a table grid.

(1) Manual Grid Creation:

- Select the table region
- (Switch to Table toolbar)
- Click on “Create empty grid” (this creates a grid with only the outer four lines)

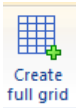
- Use the “Add … line” tools to manually add as many grid lines as required



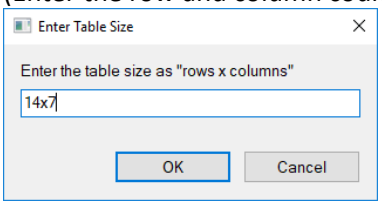
- Use “Edit grid points” for refinement (move junction points and/or grid lines)

(2) Full Grid Creation:

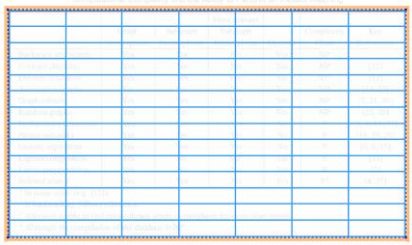
- (Specify row and column count in the table attributes)
- Select the table region
- (Switch to Table toolbar)
- Click on “Create full grid”



- (Enter the row and column count if prompted)



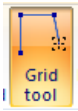
- This creates a symmetric grid:



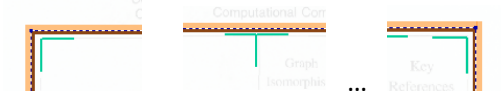
- Use “Edit grid points” for refinement (move junction points and/or grid lines)

(3) Assisted Grid Creation:

- Select the table regions
- (Switch to Table toolbar)
- Activate “Grid tool”



- Click on corner and outer junction points in clockwise order, starting from top left corner

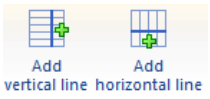


- (The cursor changes from a junction to a corner point when you are close to a table corner. Use the CTRL key to force the next corner point)
- Once all outer points are defined, the tool switches to adding grid lines (use CTRL to switch from horizontal to vertical lines)
- Press ESC to finish

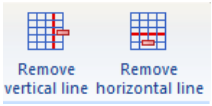
To modify an existing grid:

- Use “Edit grid points” to move junction points and/or vertical/horizontal lines

- Use the “Add … line” tools add grid lines (new lines mimic the shape of the closest existing line)



- Use “Remove … line” tools to remove selected lines (point & click mouse to remove)



- Use “Remove grid” to remove the whole table grid from a selected table region

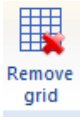


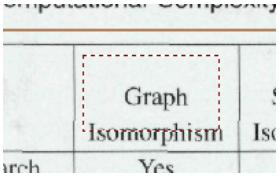
Table Cells

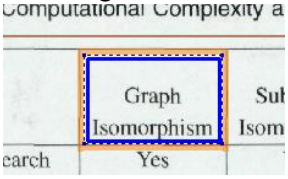
Table cells are nested regions with a table.

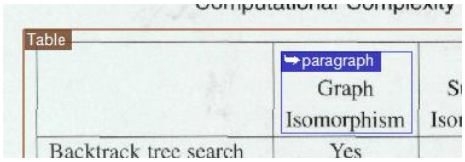
There are three ways to add cell regions:

(1) Add cell regions manually:

- Switch to Regions toolbar tab
- Use region tools to create new regions inside a table region (e.g. rectangle tool)

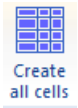






(2) Create cell region for grid cells

- Select a table region
- (Switch to Table toolbar tab)
- (Create grid)
- (Adjust the padding value – added space between grid lines and cell regions)
- Click on “Create all cells”



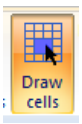
- Delete and or merge cell regions as required (use tools in Region toolbar)

	Graph Isomorphism	Subgraph Isomorphism	Error-tolerant Subgraph Isomorphism	Optimal	Complexity Class	Key References
Backtrack tree search	Yes	Yes	No	Yes	NP	
Forward checking	Yes	Yes	No	Yes	NP	[32]
Discrete relaxation	Yes	Yes	Yes[1]	Yes	NP[2]	[12]
Association graphs	Yes	Yes	No	Yes	NP	[14, 23]
Graph edition	Yes	Yes	Yes	Yes	NP	[7, 21, 36]
Random graphs	Yes	Yes	Yes	Yes	NP	[25, 38]
Probabilistic relaxation	Yes	Yes	Yes	No	P	[5, 8, 11, 37]
Neural networks	Yes	Yes	Yes	No	P	[16, 29, 28]
Genetic algorithms	Yes	Yes	Yes	No	P	[6, 9, 15]
Eigendecomposition	Yes	No	No[3]	Yes	P	[33]
Linear programming	Yes	No	No	Yes	P	[2]
Indexed search	Yes	Yes	No	Yes	P[4]	[4, 27]

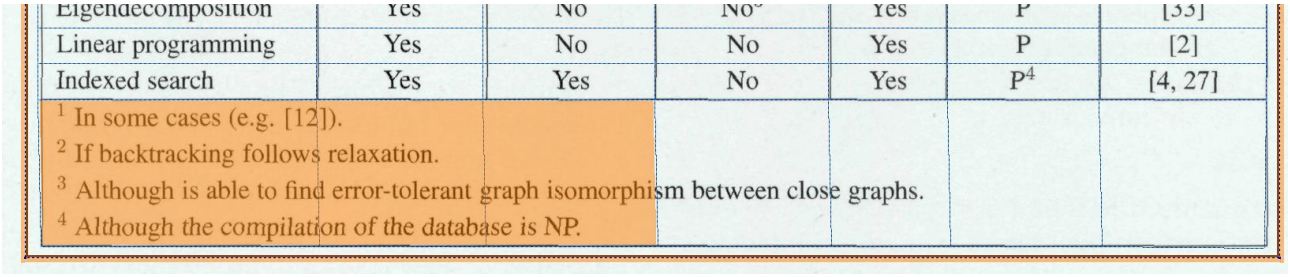
[1] In some cases (e.g. [12]).
[2] If backtracking follows relaxation.
[3] Although is able to find error-tolerant graph isomorphism between close graphs.
[4] Although the compilation of the database is NP.

(3) Use cell drawing tool:

- Select a table region
- (Switch to Table toolbar tab)
- (Create grid)
- (Adjust the padding value – added space between grid lines and cell regions)
- Activate the “Draw cells” tool

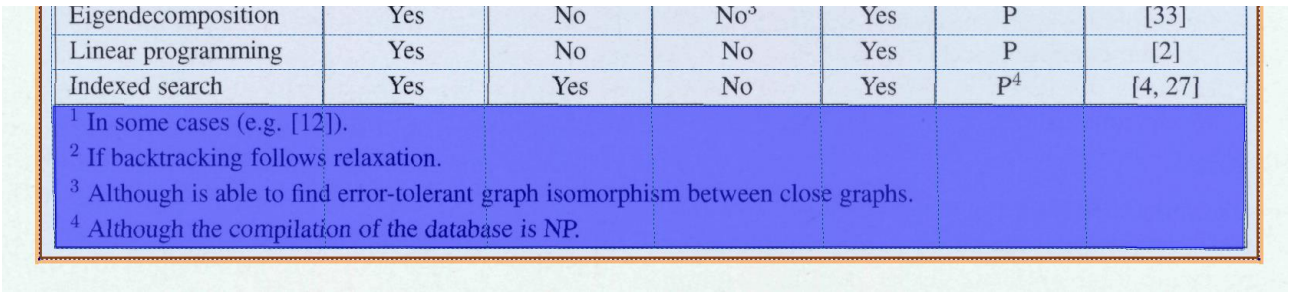


- Click on a grid cell to create a small cell region or click and drag the to create larger cell regions (spanning multiple rows / columns)



Eigendecomposition	Yes	No	No[3]	Yes	P	[33]
Linear programming	Yes	No	No	Yes	P	[2]
Indexed search	Yes	Yes	No	Yes	P^4	[4, 27]

[1] In some cases (e.g. [12]).
[2] If backtracking follows relaxation.
[3] Although is able to find error-tolerant graph isomorphism between close graphs.
[4] Although the compilation of the database is NP.

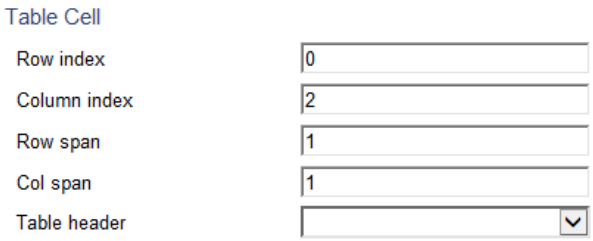


Eigendecomposition	Yes	No	No[3]	Yes	P	[33]
Linear programming	Yes	No	No	Yes	P	[2]
Indexed search	Yes	Yes	No	Yes	P^4	[4, 27]

[1] In some cases (e.g. [12]).
[2] If backtracking follows relaxation.
[3] Although is able to find error-tolerant graph isomorphism between close graphs.
[4] Although the compilation of the database is NP.

Cell Attributes

Cell regions have specialised attributes. They appear together with the usual region attributes in the respective dialog.



Additional Cell Tools

Use “Select all cells” to select all cell regions of a table.

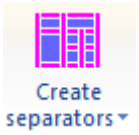
Use “Recalculate cell positions” to fill in the row/column index and row/column span attributes for al cells as seen above.

Use “Add lines to cells” to create a text line object for each table cell region (this is equivalent to using the “Initial line” tool in the Text Line toolbar).

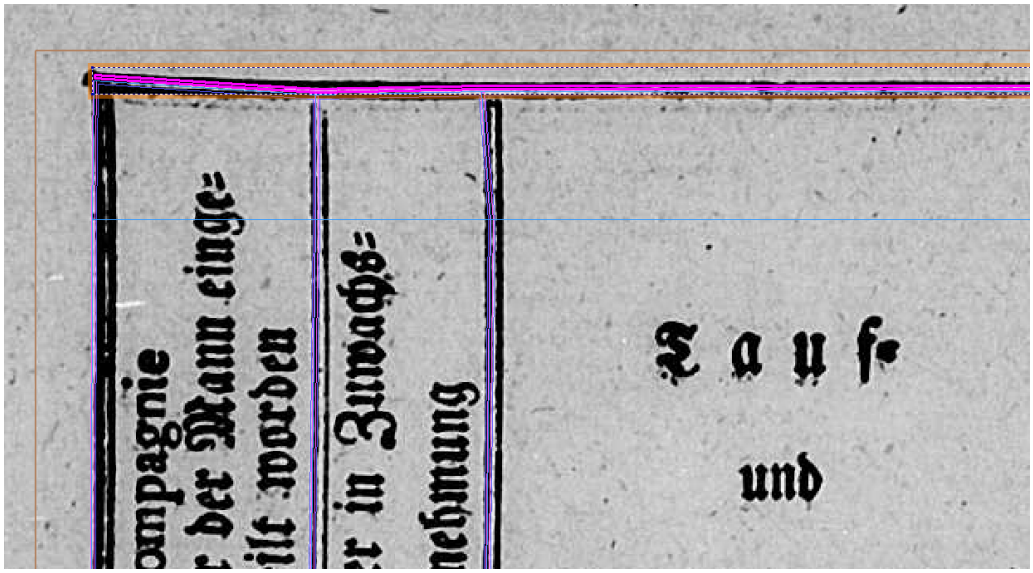
Table Separators

For explicit separator regions in addition to the grid:

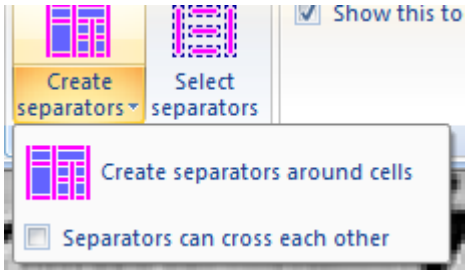
- Use the “Create separators” tool

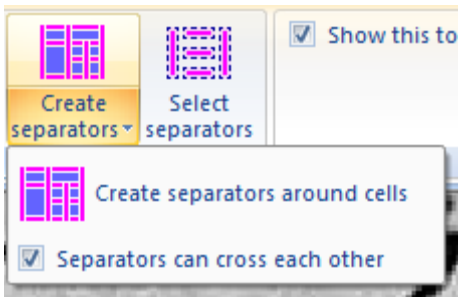


- This adds separators around cells

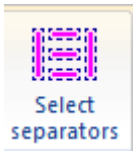


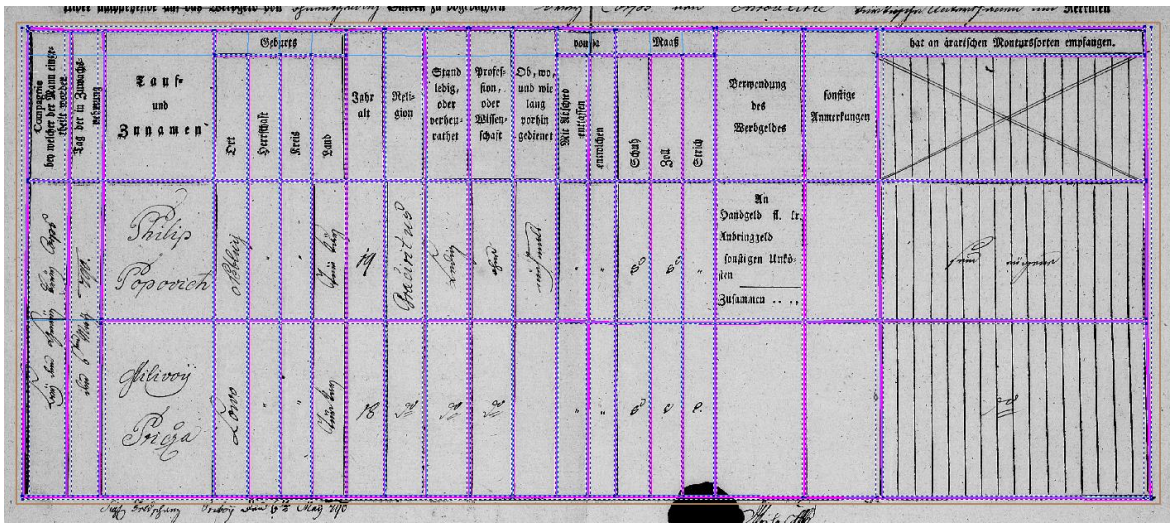
- Vertical separators are created first
- There is an option to create either short horizontal separators that do not cross the vertical separators or to create long horizontal separators that do cross the vertical separators





- Use “Select separators” to select all separator regions of the table





Page Object Properties and Text Content

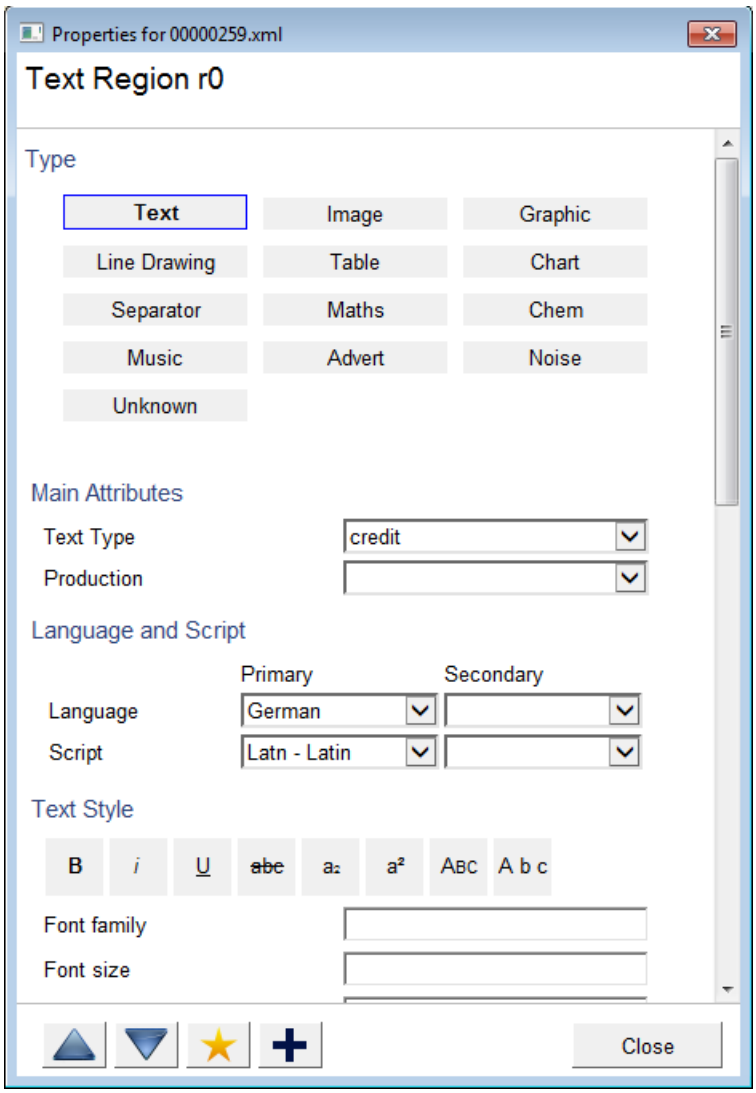
Properties

Each layout object has properties (or attributes). Depending on the type (text, image, word, ...) there are different properties available.

To specify the properties for an object:

- Click '...Data' in the toolbar panel called 'Properties' (keyboard shortcut F10) or double click the object

A dialog opens:



The dialog contains all properties the selected object. The heading shows the type and the ID of the current object. The ID is generated internally and cannot be changed.

There are 5 different types of properties:

- List properties – displayed as a drop-down box with all available items (e.g. 'Text colour': 'black', 'red', ...)
- True/False properties – displayed as a drop-down box with the items 'True' and 'False' (e.g. 'Indented')
- Number properties – displayed as a text input field (e.g. 'Font size')
- Text properties – displayed as a text box (e.g. 'Plain text')
- Font style properties – displayed as buttons

Furthermore each property can have the 'Not set' state, which means they have not been set yet or have been reset. To set this state for a property either select the empty item from the drop-down box (for list and true/false properties) or clear the text field (for number and text properties).

Changes take effect immediately and can be reverted using the undo functionality.

To navigate to another region click the arrow buttons.

Updating Multiple Objects

It is possible to change properties of multiple objects at once. To do so select the desired objects and open the properties dialog if it is not already open.

Changing a property now changes it for each selected region.

If the regions have different values for a property, this is indicated by three points within the dialog:

Saving the Current Properties as Default

It is possible to change the default properties for a region type. Each time a new region is created, it will be initialised with these default values.

To save the current property setting as default, click the star button.

For more information on settings and how to restore the original settings, see the chapter 'Customisation'.

Custom Attributes

It is possible to add more object attributes.

To add user-defined page attributes:

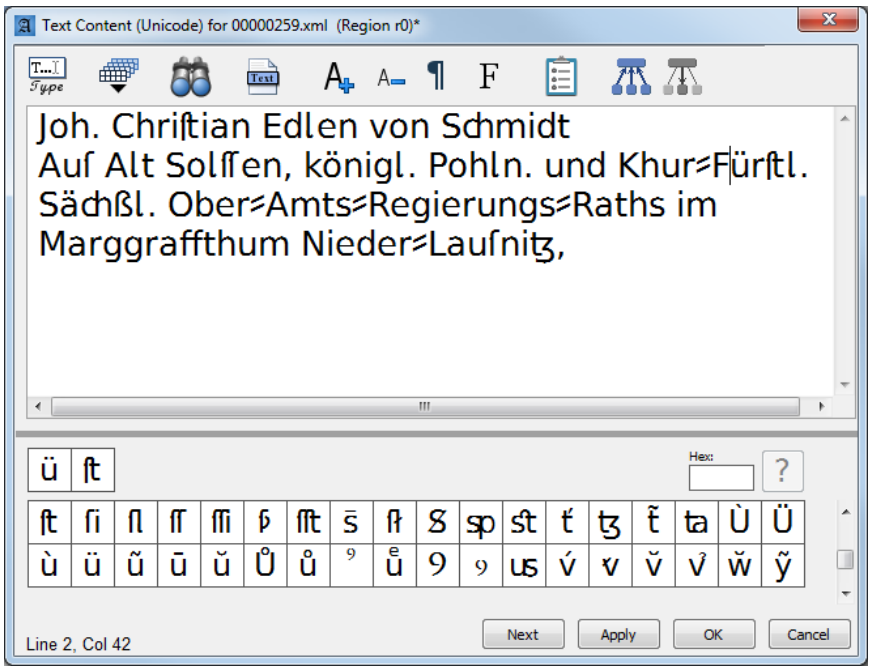
- Click on the "+" button at the bottom

See the Page Attributes section for more details.

Text Input Dialog

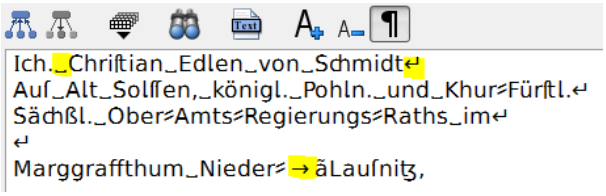
The text content for regions can be input using a special dialog. To open it, click ‘Text Content’ in the toolbar panel called ‘Text’ (keyboard shortcut F11).

The dialog also offers a virtual keyboard with special characters and an input field to manually enter special characters using their hexadecimal code number.



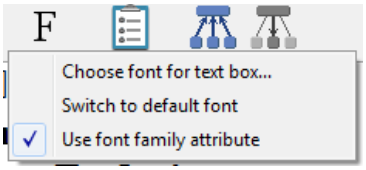
For better readability the outline of selected objects won’t be visible while the text input dialog is open. Instead the objects are highlighted with a transparent overlay.

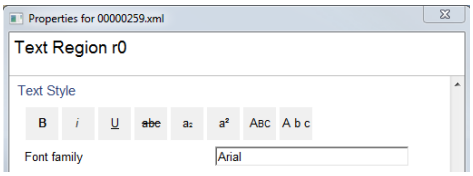
The font size of the text field can be adjusted using  . Whitespaces can be visualised using the  button.



Use the  button to select a different font for page text content. This font will be used within the text dialog, for text overlay in the main view, and for the virtual keyboard. The default is “Aletheia Sans”. You can also enable an option to use the “Font family” attribute (from the properties dialog) as display font in the

text dialog.

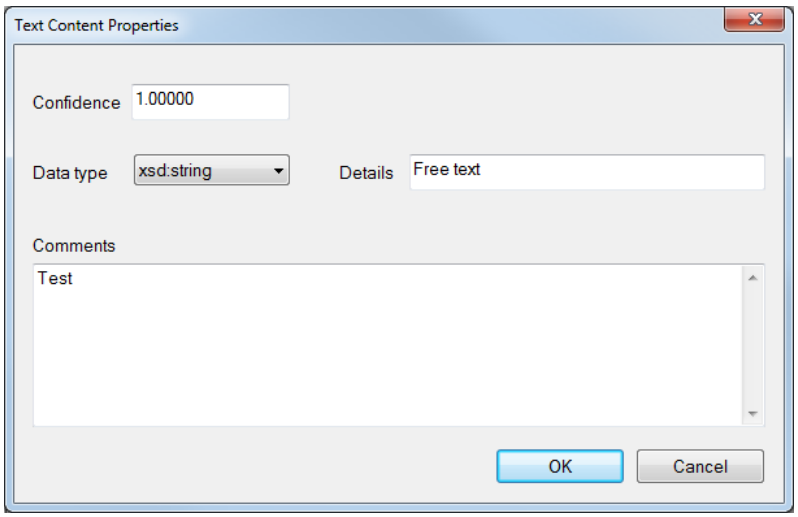




Note: When working with languages that use non-Latin scripts (e.g. Arabic or Chinese), changing the font to something more suitable to that language will improve how text is displayed. Times New Roman works well for Arabic, for example.

Additional properties related to the text content can be viewed and edited using the  button.
This opens a dialog with the following fields:

- Confidence: OCR confidence or similar (values from 0.0 to 1.0)
- Data type: Observed or expected data type of the text content
- Data type details: Refinement of the data type field. Can be a regular expression, for instance
- Comments: Generic comments on the text content (e.g. producer)



Composing and Extracting Text

If text is already specified for child and/or parent objects, it is possible to use that text to fill the text of the current object.

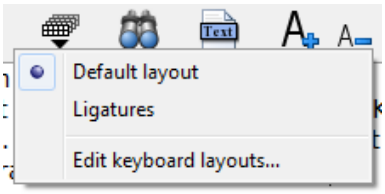
To compose the text using the child objects, click  . If for instance the current object is a text line, the text would be composed using the words assigned to the line.

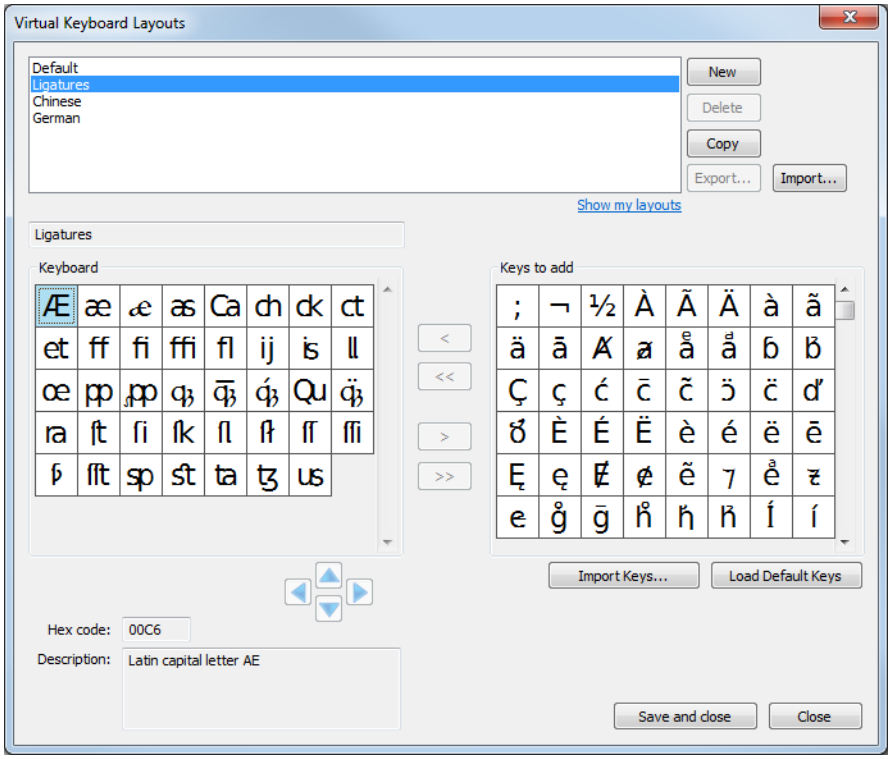
To extract the text from the parent object, click  . If for instance the current object is a text line, the text would be extracted from the parent text region.

To propagate text across several layers (e.g. from regions to lines to words), use the tool 'Propagate text recursively' (from the 'Tools' menu). This will compose the text for all children and children's children of the selected regions.

Virtual Keyboard Layouts

It is possible to modify, create and use different layouts for the virtual keyboard. Press  to choose a layout or open the editor:





Selecting a Layout

Double click a layout in the list to select it and close the editor. The selected layout will be displayed in the text dialog.

Creating a Layout

To create a new layout either click ‘New’ or select an existing layout and click ‘Copy’ to create an identical copy. Then change the name of the layout (text field beneath the list).

Modifying a Layout

Only private layouts can be modified. All layouts already included in Aletheia are read-only (but can be copied). The pool of available character keys is displayed on the right of the dialog (“Keys to add”). Double click a key or select it and click ‘<’ to add it to the current layout. Double click it again or click ‘>’ to remove it. The position of a key can be changed by selecting it and moving it by using the arrow buttons.
All changes will be saved when clicking ‘Save and close’.

Importing Keys

Keys can be imported from already transcribed pages (in PAGE XML format).
To import keys:

- Click “Import keys…” (under “Keys to add”)

- Select one or multiple PAGE XML files (ALTO and FineReader XML will work as well)
- The keys will appear sorted by frequency (keys already on your keyboard will not be shown)
- Use the "<" or "<<" buttons to add the keys to your keyboard
- Optional: select each key in your keyboard and enter a description

Click "Load Default Keys" to reset the list of available keys to the default list.

Exporting / Importing Layouts

Use "Export…" and "Import…" to save/load keyboard layout XML files. Note that the export will only work on private layouts (layouts that have been previously created or imported).

Alternatively, private layouts can be exported by clicking the link 'Show my layouts'. This will open the folder for the local layout files. By copying the files to the folder <Aletheia>/bin/data/keyboard_layouts of an Aletheia copy, the layouts will appear in the keyboard layout editor as 'read only' layouts.

On-Image Transcription

The transcription tool positions a text input box directly above the text on the image. Keyboard shortcuts allow you to keep transcribing the whole page via the keyboard alone.

To activate the transcription tool:

- Click the  icon within the text dialog
- Select a text region (or text line object, word object etc.)
- A dialog is displayed on the image

To enter or correct text:

- Type the text as seen in the image
  - For special characters use the keyboard in the text dialog (the text in the transcription dialog is synchronised with the text in the text dialog and vice versa)
- For regions: Press ENTER to move to the next text line
  - Press ENTER with blank text to move to the next region
- Press PAGE DOWN to move to the next text object
- Use the cursor keys up/down to manually reposition the text input box
- Use the mouse or SHIFT + cursor keys to pan the image as required
- Adjust the font size using the toolbar in the text dialog

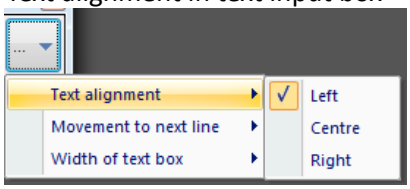
Keyboard shortcuts:

- Page up / page down: Previous / next text object
- Cursor up / down: Move text input box up / down

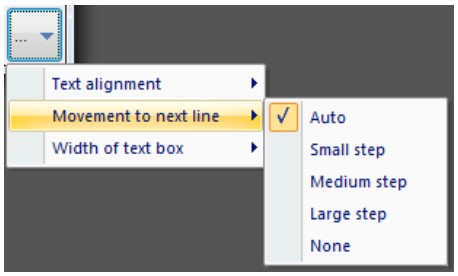
- Cursor left / right: Move text cursor within text box
- Shift + cursor keys: Pan image
- Ctrl + cursor keys: Select neighbour object
- Enter:
    - When working on text regions: Next text line (or next region, if blank text)
    - When working on text line objects, word objects or glyph objects: Next object
- Shift + Backspace: Previous text line (when working on text regions)

Options:

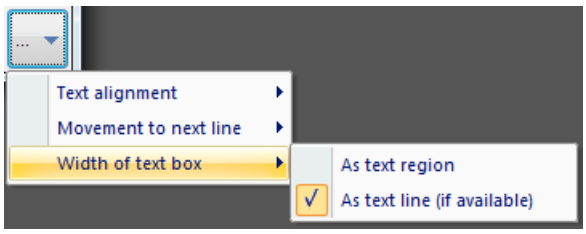
- Text alignment in text input box

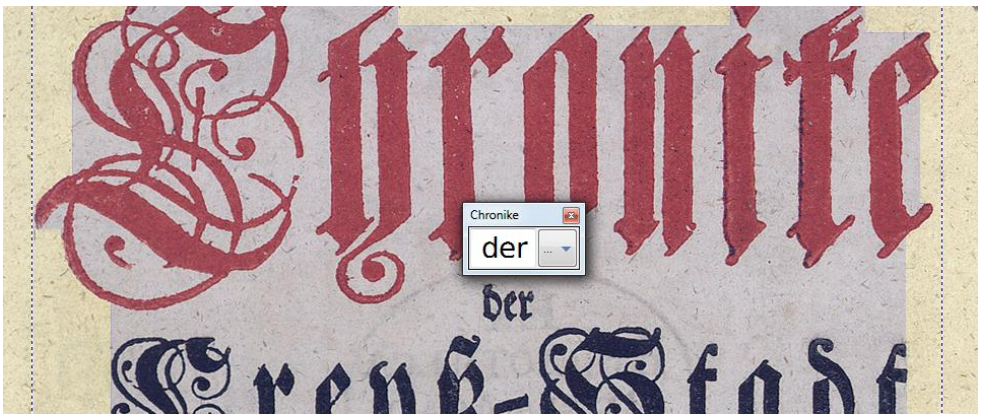


- Behaviour when pressing ENTER to move to next text line:
    - Auto: Uses the position text line objects (outlines) to move the text input box
    - Small, medium or large step: Moves a fixed distance down
    - None: Does not move the input box – use the cursor keys to reposition manually



- Width of text input box:
    - As text region: The text box always has the same width as the text region, even if the text line is short
    - As text line: The text box has the width of the current text line, if this information is available (text line outline)





Text Overlay

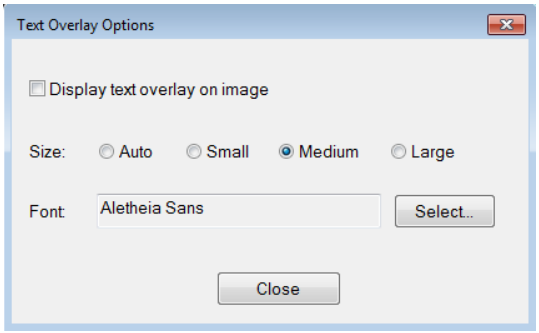
For checking the text content of regions (and text lines, words or glyphs) a text overlay can be activated. This will display the text of regions as semi-transparent layer on the document image.

To activate the overlay:

- Tick the checkbox ‘Text Overlay’ in the toolbar panel called ‘Text’ (keyboard shortcut Ctrl+T)

OR

- Click the  icon within the text dialog
- Enable the text overlay checkbox and/or adjust settings

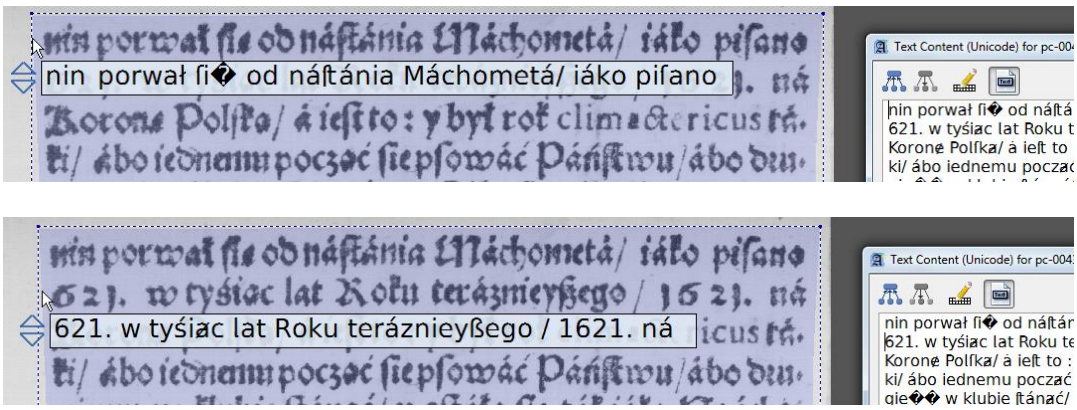


Text Overlay for Regions

If no region is selected, the text of the region at the current position of the mouse is displayed:

Once a region is selected, only the text of this region is displayed next to the mouse cursor.

If the text dialog is open, the overlay shows the text of the currently ‘focused’ line of the text input field (the line where the cursor is). That way, the text can be quickly checked line by line using the cursor keys:



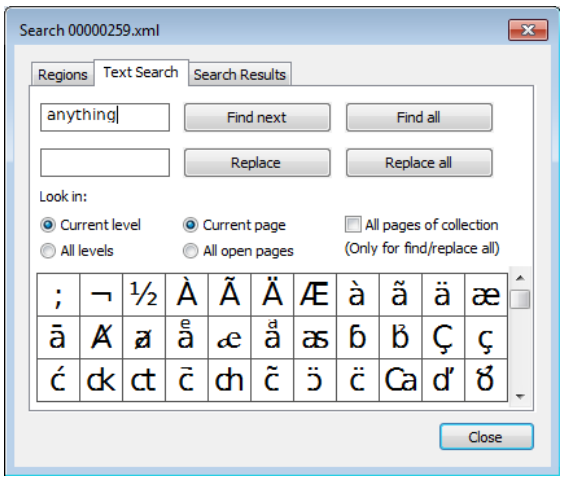
Text Overlay for Lines, Words, and Glyphs

The text for lines, words and glyphs is displayed above the objects, either when hovering over them with the mouse or when selecting them:

Text Search and Replace

To do text search or replace:

- Press Ctrl + F
  OR
- Click on “Find…” in the Edit menu
  OR
- Click on the search icon in the text dialog
  OR
- Open the “All Regions” dialog and switch to the search tab
  

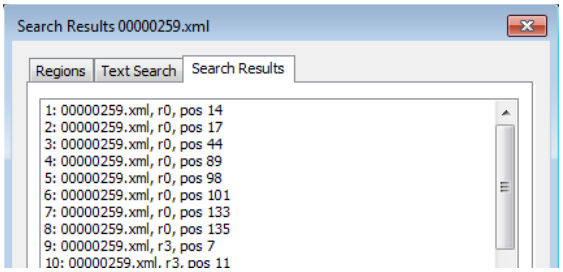


To find individual text occurrences:

- Enter a search term next to the “Find next” button
- Click on “Find next”. This will open the text dialog and highlight the found text or show a message that nothing was found.

To find all text occurrences:

- Enter a search term
- Click on “Find all”. This will show the search results in a separate tab.



- Click on a result item in the list to highlight it in the text dialog

To replace individual text occurrences:

- Enter a search term
- Enter a replacement
- Click on “Replace”. If nothing is highlighted yet, this will only highlight the first match.
- Click again on “Replace” to replace the found text and jump to the next occurrence

To replace all text occurrences:

- Enter search term and replacement
- Click on “Replace all”
- Confirm the message that pops up

Options:

In Aletheia and PAGE XML, text can be stored in different page object levels: regions, text line objects, word objects, and glyph objects. Use the search options to select where to search.

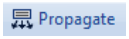
- Level:
  - Current: Objects you are working on right now (either regions, text lines, words or glyphs)
  - All: Search across all the above objects
- Page(s):
  - Current page: Only search in the page that is in the foreground
  - All open pages: Search in all pages that are currently open in Aletheia

    - All pages of collection: This includes all pages that appear in the page collection pane, even the ones that are not open. This option only applies to “Find all” and “Replace all”.
- Case-sensitive search:
    - Tick this option to match upper and lower case during text search and replace

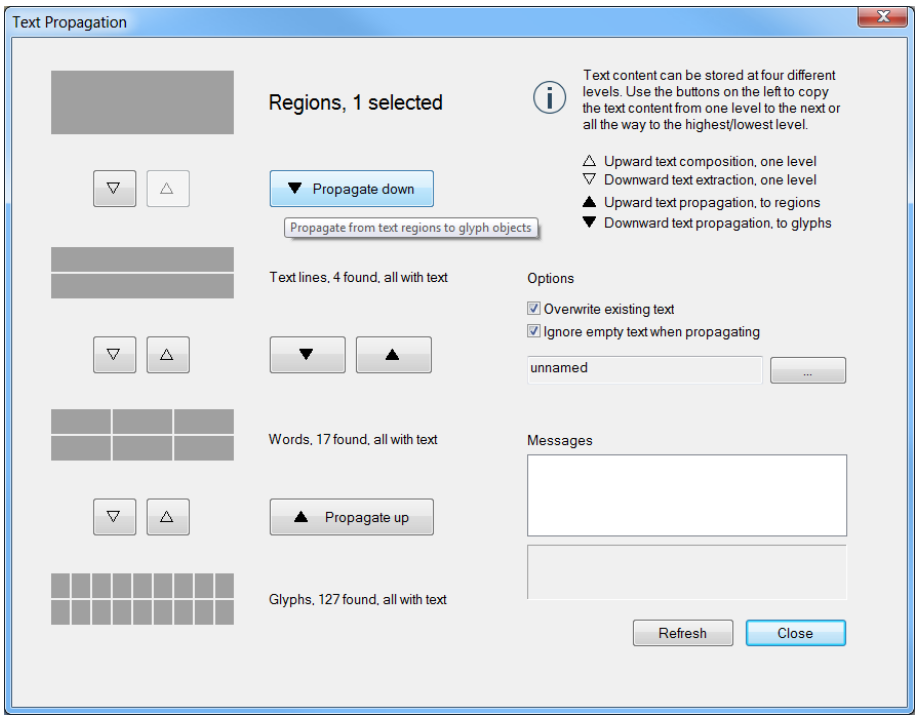
Text Propagation

Text content can be stored at four different levels: In region objects, in text line objects, in word objects and in glyph objects. The text propagation dialog allows to copy the text content from one level to the next or all the way to the highest/lowest level.

To propagate text:

- Select objects (regions, text lines, words or glyphs)
- Click on “Propagate” 

(opens dialog)
- Click on the appropriate propagation button for the direction and depth you want to propagate

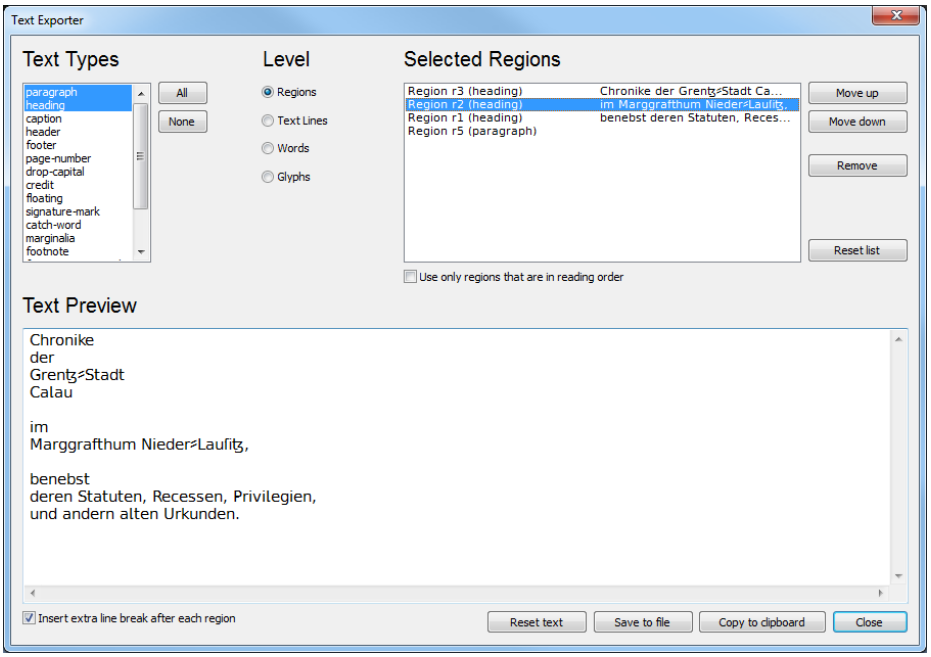


If the structure of your page layout does not match the text content that is to be propagated, error messages will be shown on the bottom right. Select a message to see details.

A custom method for text propagation can be set up using the “…” button. For more details see “External Text Propagation”.

Text Export

All text of a document can be serialised and exported to a text file. The text export dialog can be opened by clicking ‘Export text…’ in the toolbar panel called ‘Text’.



The regions contributing to the final text can be filtered by type. The regions are ordered using the reading order (for regions that are within the reading order) and y-position on the page (for regions not in reading order). However, the resulting list of regions can be edited manually as well (move regions, delete regions).

The ‘Level’ determines the source objects of the text (text regions, text lines, words or glyphs). It is however not possible to export text from specific text lines, words or glyphs. Filtering is only available on region level.

The final text is displayed at the bottom of the dialog and can be edited if needed. It can then be saved as a text file or copied to the Windows clipboard.

The text content of a selected region can also be quickly exported via the Windows clipboard:

- Select a text region
- Press CTRL + C
- Paste in another Windows application with CTRL + V

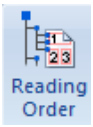
Reading Order and Structure

3.4 ↔4.0

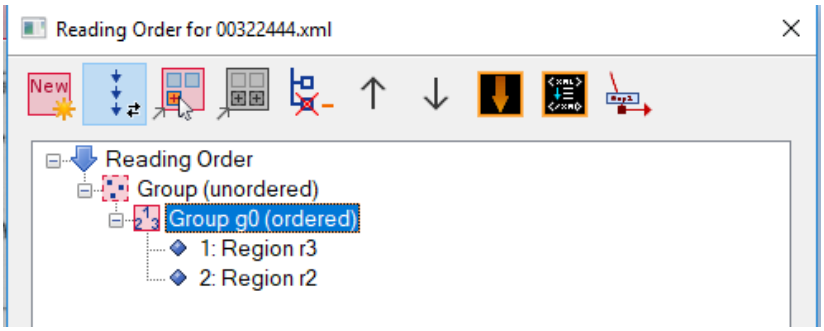
The reading order specifies in which order to read text regions. In Aletheia 4, the concept was extended to cover more of the page structure in general. A new “Structure” toolbar tab and view were introduced.

For text line order see page 54.

To view or define the reading order click ‘Reading Order’ in the Regions or Structure toolbar (keyboard shortcut F12).



The view will change and a dialog will pop up.



The reading order is a hierarchy of groups and references to regions. A group can contain sub groups as well as region references. A group can be ordered (default) or unordered. An unordered group can be used if some regions are related in a sense but there is no specific order in which to read them (e.g. advertisements).

Creating groups

By default there is already an ordered base group.
To add more groups:

- Select the parent group within the reading order tree
- Click the ‘Add new Group’ button of the tool bar

A new group is always ordered. To change it to an unordered group, click the ‘Toggle Order Status’ toolbar button.

Adding regions to a group

There are two ways to add regions to a group:

1. “Add to group” tool
    - Select the group to which you want to add regions
    - Activate the “Add to group” tool 
    - Click on the regions in the desired order (in the image view)
2. By selecting regions
    - Switch to the Region toolbar tab
    - Select the group to which you want to add the references (in the reading order dialog)
    - Select all regions you want to add (in the image view)
    - Click the ‘Add’ toolbar button 

    Note: The regions will be added in vertical order (top-to-bottom).

Moving elements

It is possible to move elements (sub groups, region references) of a group up and down to change the order. To do so use the ‘Move up’ and ‘Move down’ toolbar buttons.

It is also possible to move elements to another group via drag and drop. Left click the element to move (do not release the mouse button) and drag the element to the desired destination group.

Deleting elements

To remove an element (sub group, region reference) from a group, select the element within the reading order tree and click the ‘Delete Element’ toolbar button (or use the ‘Delete’ key).

Automatically creating reading order

To automatically create the reading order containing all text regions sorted vertically:

- Click the ‘Create top-to-bottom reading order’ button

To automatically create the reading order containing all text regions as they are ordered internally (as in the source XML file or in the order the regions were created):

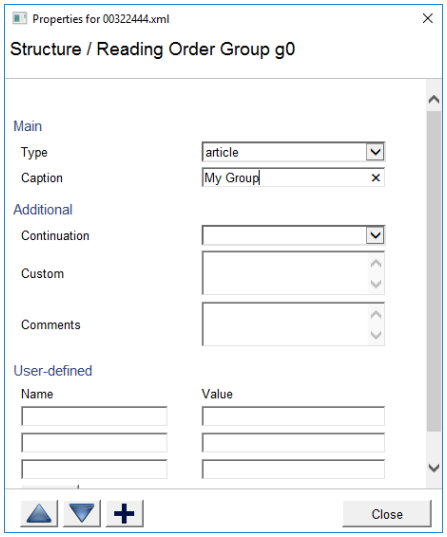
- Click the ‘Create from internal list’ button

Group Attributes

To modify group attributes:

- Open the attributes dialog

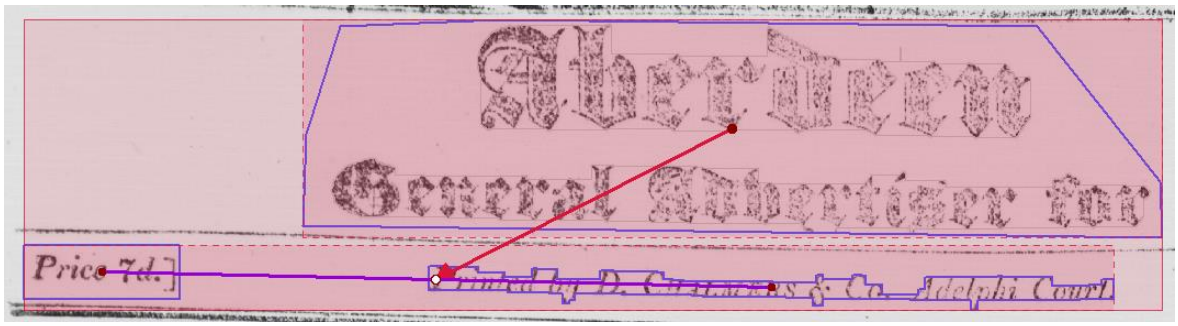
- Select a group
- A dialog opens:



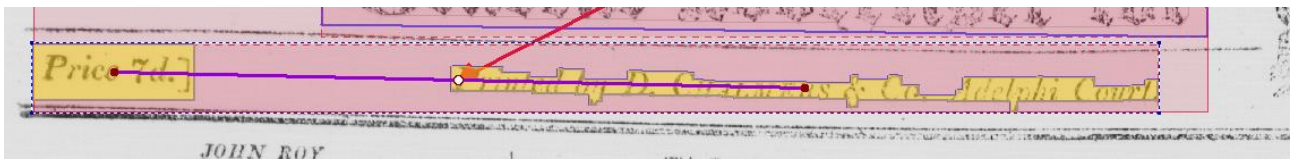
Viewing the reading order

Once the reading order is defined, it will also be displayed within the document image. Ordered groups are represented by arrows connecting their elements in the specified order. Unordered groups are represented by star like lines, reaching from the group centre to the child elements of the group. Sub groups will be displayed with another colour than their parent groups.

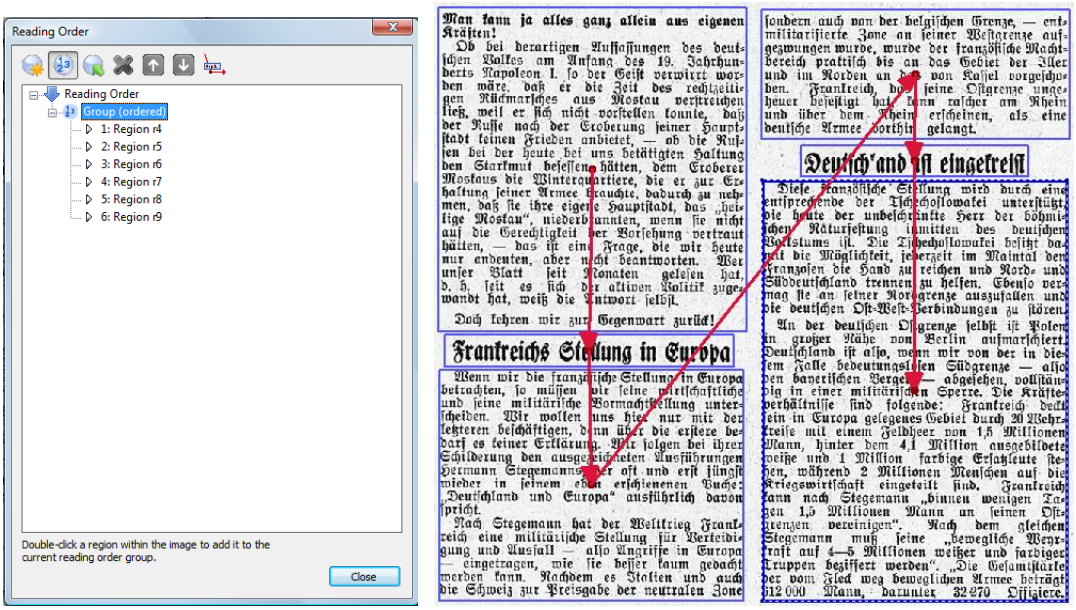
When in “Structure” view, groups will also be outlined by a red semi-transparent rectangle. Un ordered groups have a dashed outline.



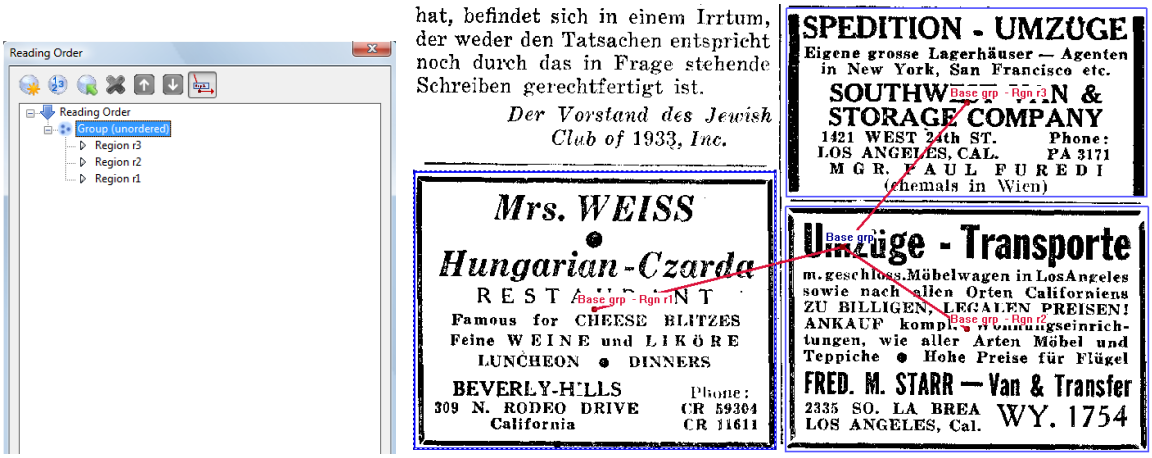
Regions in selected groups are highlighted:



Example of an ordered group:

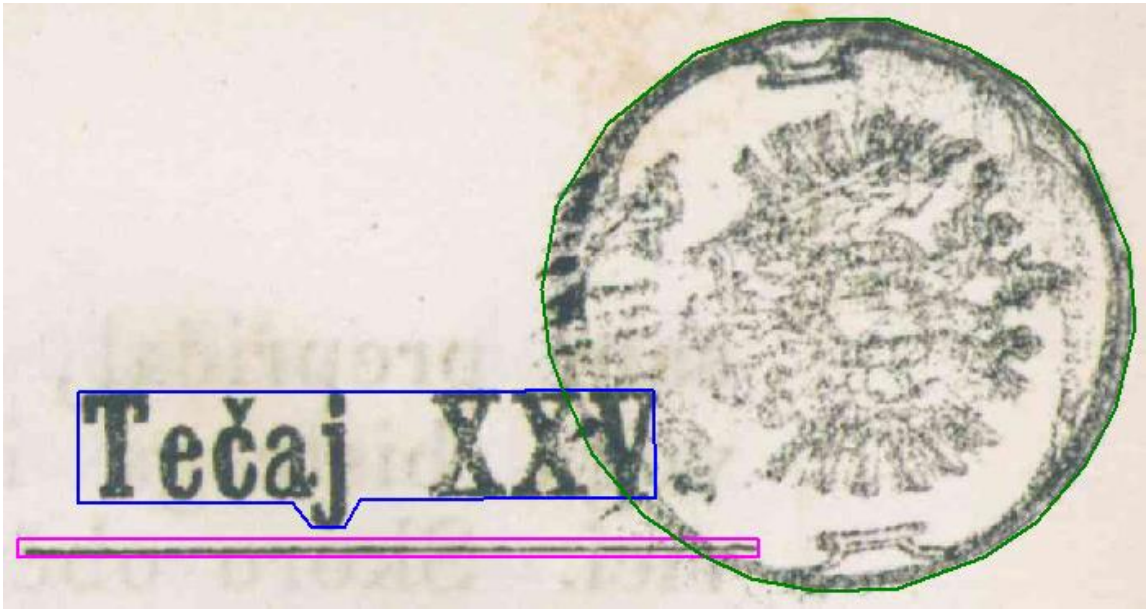


Example of an unordered group:

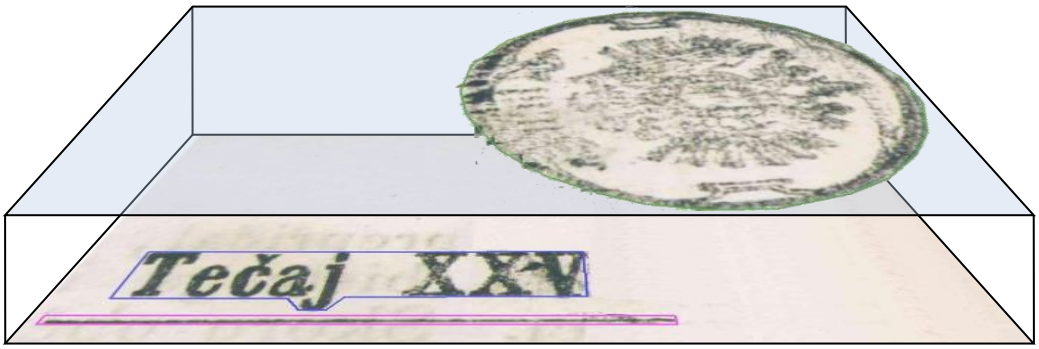


Layers

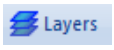
Layers are used to group regions belonging to the same logical document level. The main purpose is to avoid invalidities and problems caused by overlapping regions. See the following example:



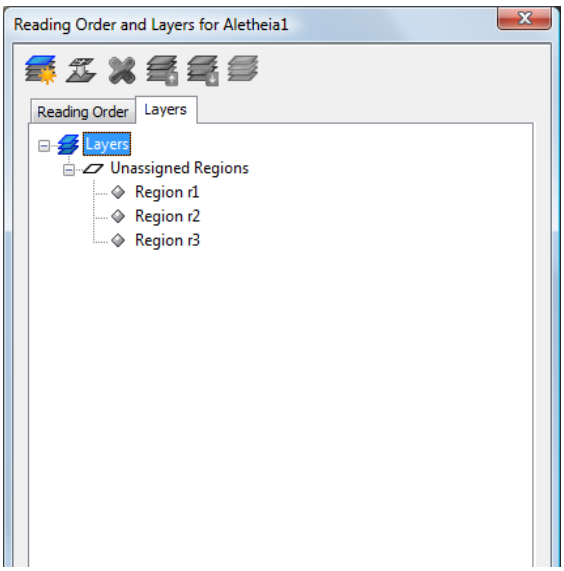
The stamp region overlaps the text region and the separator region. We can now define two layers, one for the text and the separator and one for the stamp. Without using layers, this would an invalid document layout (by definition).
The layers can be imagined as transparent sheets stacked one over another (see illustration). In this example the stamp would naturally be the topmost layer (or front layer).



To manage layers in Aletheia, open the ‘Layers’ dialog by clicking ‘Layers’ in the toolbar panel called ‘Structure’.



Shared dialog for layers and reading order:

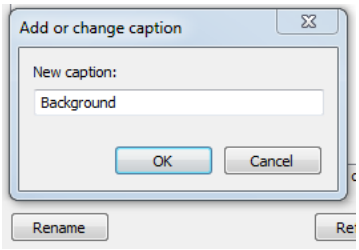


Creating Layers

Click the 'Create Layer' toolbar button. The new layer appears as new item at the top of the tree.

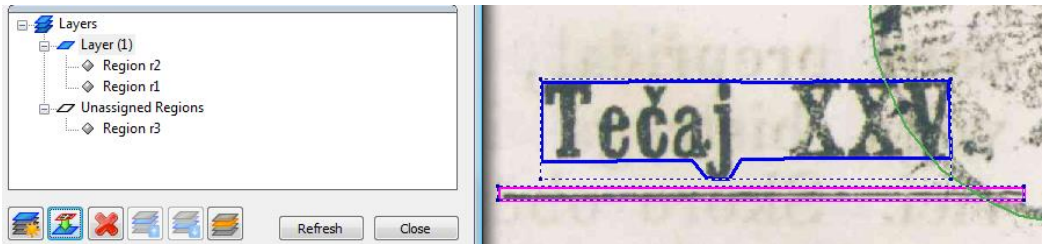
To add a caption:

- Select the layer and click 'Rename'



Assigning Regions to a Layer

Select the regions you want to assign (in the main document view), then select the layer you want them assign to (in the layer dialog) and press the 'Assign' icon. The regions will appear as children of the layer. It is also possible to select a layer in the tree and add regions by double clicking them in the document.



Removing Regions from a Layer

Select the region you want to remove (in the tree) and click the ‘Remove’ icon or press the delete key.

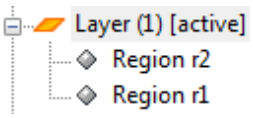
Moving Layers to the Front or Back

Layers can be moved in the theoretical third dimension. To move a layer in direction of the front, press the ‘Move up’ icon. To move a layer in direction back, click the ‘Move down’ icon.

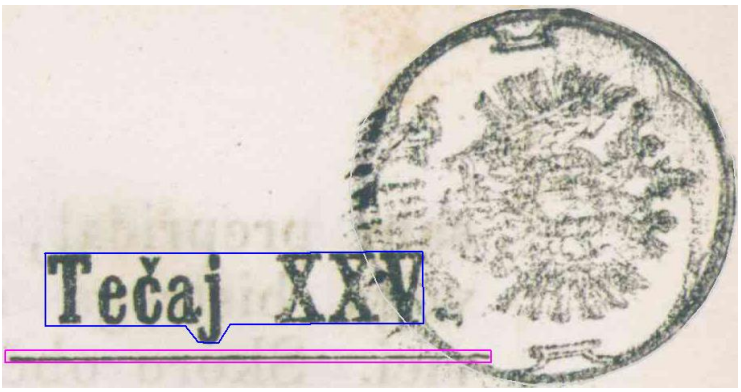
Activating a Layer

To activate a layer, click the ‘Activate’ icon.

The active layer will be highlighted in the tree:



Regions not belonging to the active layer will be hidden in the document view:



Regions that are created while a layer is active will be assigned to this layer automatically.

Note: A layer stays only active while the layer dialog is open.

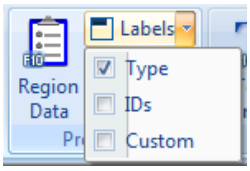
Deleting a Layer

To delete a layer select it within the tree and click the ‘Delete’ icon or press the delete key.

Additional Viewing Options

Displaying Region (Sub)Types, IDs and the Custom Attribute

By default the type of a region is displayed in the top left corner of its outline. If the region has a sub-type this will be displayed instead of the main type. The type labels can be hidden by deactivating 'Labels-Type' in the toolbar panel called 'Properties'.



For text lines, words and glyphs only ID and Custom labels are supported.

For the custom label the content of the attribute called 'custom' is used (see "Properties").

Bounding Boxes of Connected Components

To view the bounding boxes of all connected components of the black-and-white document image, activate 'Bounding boxes' in the toolbar panel called 'Image' (in the View toolbar tab).

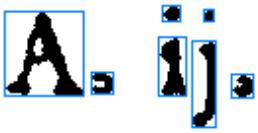
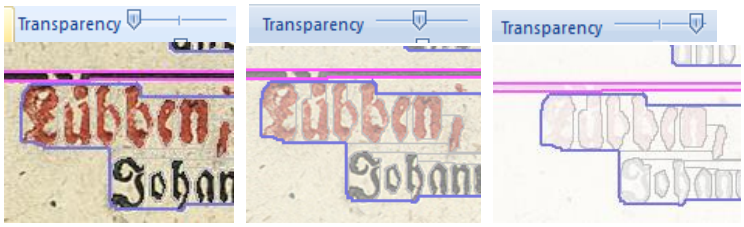


Image Transparency, Brightness and Contrast

These settings are temporary view enhancements and are not permanently applied to the image(s).

To select the transparency of the document image:

- Go to the 'View' toolbar
- Use the Transparency slider



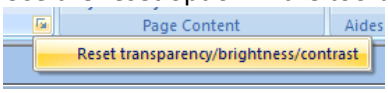
To change brightness and contrast (a balance of both might deliver best results):

- Go to the 'View' toolbar

- Use the sliders

To reset the settings:

- Use the reset option in the toolbar menu



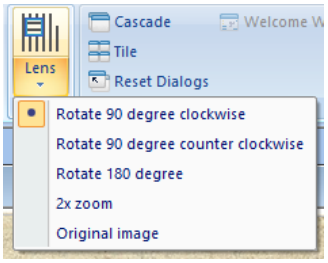
View Lens

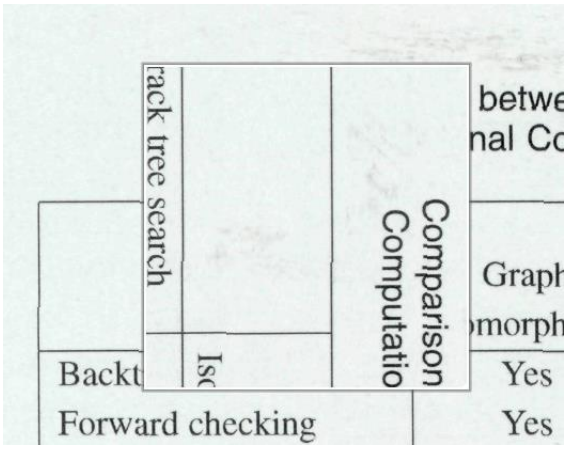
A movable and resizable lens with different effects can be used for better readability.

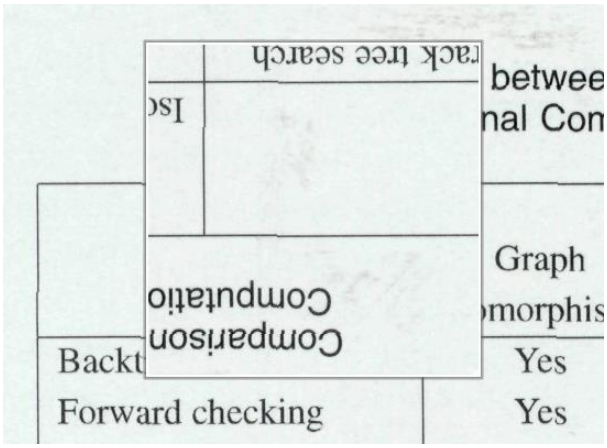
To use the lens:

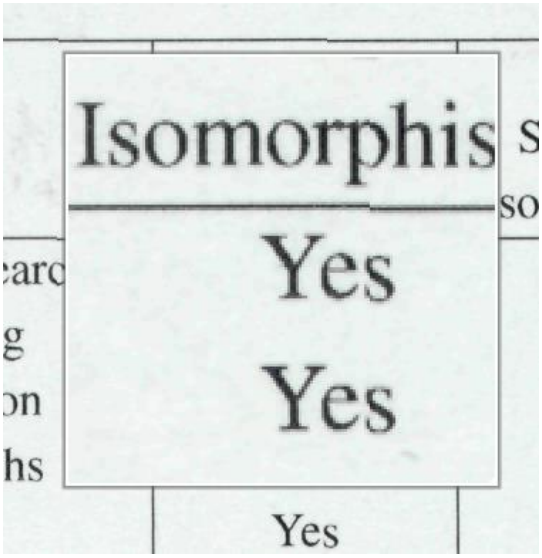
- Switch to View toolbar tab
- Click on “Lens”

- (Position and resize as required)
- Right click on the lens to switch between different effects (rotation, zoom, …)
  OR
- Select a effect from the drop-down menu



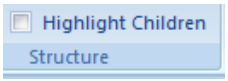






Highlighting Child Regions

To check which sub-region (text line, word, glyph) belongs to which parent region (text block, text line, word), the relationship between the regions can be displayed. To do so, activate ‘Highlight Children’ in the toolbar panel called ‘Structure’.



Highlighting in ‘Region’ mode:

Highlighting in ‘Text Line’ mode:

Highlighting in ‘Word’ mode:

Highlighting Parent Regions

The relation between parent and child regions can be displayed. Activate ‘Highlight Parent’ in the toolbar panel called ‘Structure’. The parent region will be highlighted as a blue overlay.

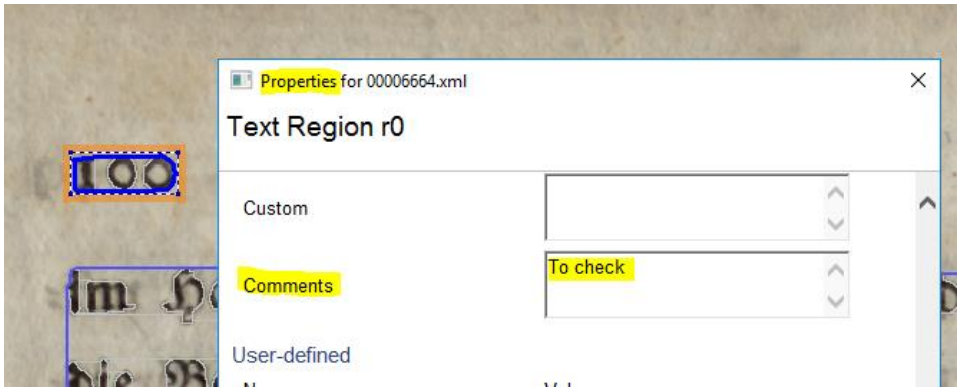
Example - Highlighting parent text line in ‘Word’ mode:

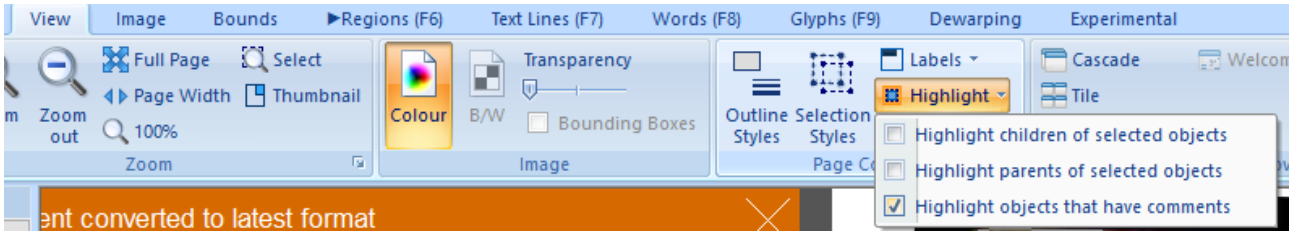
Highlighting Objects with Comments

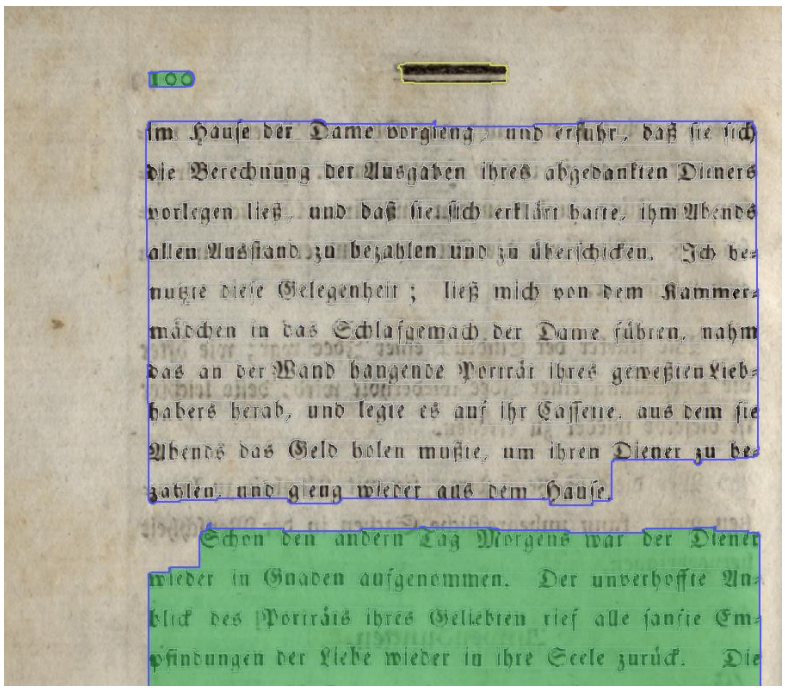
Page layout objects that have comments can be highlighted. To do so:

- (Add comments in object properties dialog)
- Go to View toolbar tab
- Enable “Highlight objects that have comments”

Objects with comments are then highlighted with a green fill colour.







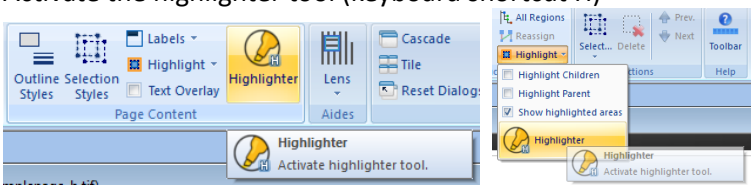
Highlighter Tool

Highlights can be added to draw attention to problems in the ground truth that need to be addressed.

Note:
Highlights are added as regions of type “Custom” and subtype “ui.highlight…”. Hence, highlights should be removed from final versions of page transcriptions in order to avoid confusion with normal page content.

To highlight areas:

- Activate the highlighter tool (keyboard shortcut H)



- Draw a polygon by
  - Left clicking to add points
  - Clicking and holding the left mouse button to draw continuously

To add comments to a highlight:

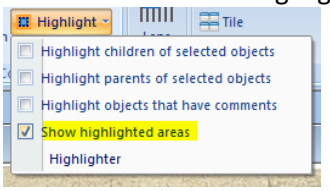
- Activate the Regions toolbar tab
- Open the Region Attributes dialog
- Select the highlight region
- Enter text under “Comments”
  OR
- Activate the highlighter tool (keyboard shortcut H)
- Double-click on a highlighted area
- Enter a comment and confirm

To view comments:

- Activate the highlighter tool (keyboard shortcut H)
- Position the mouse cursor over a highlighted area

To show/hide highlights:

- Tick or untick “Show highlighted areas”



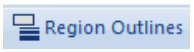
To delete a highlight:

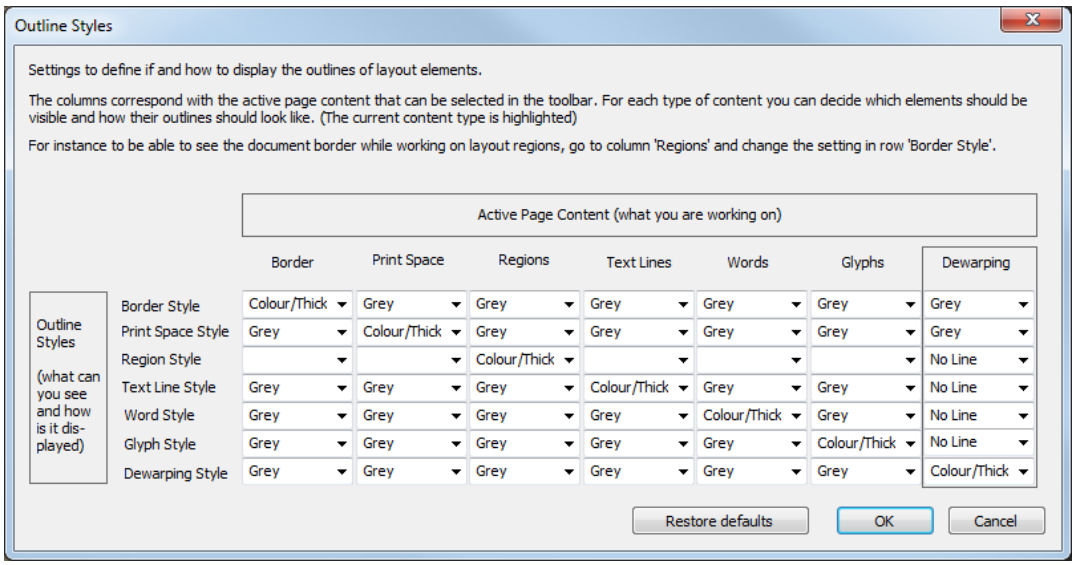
- Activate the highlighter tool (keyboard shortcut H)
- Right-click on the highlighted area and confirm
  OR
- Activate the Regions toolbar tab
- Select the highlight region
- Press the delete key and confirm

Customising Outline Styles

The outlines of layout elements (border, print space, regions, text lines, words and glyphs) are by default displayed as a thick coloured line, when in the corresponding display mode and they are displayed as a thin grey line, when in another display mode. That means if you view text lines, the lines are highlighted with a coloured line and all other regions as well the border and the print space are also visible as grey outlines.

To customise the outline styles switch to the “View” toolbar tab and click “Region Outlines” in the toolbar panel called “Page content”. A dialog will open:





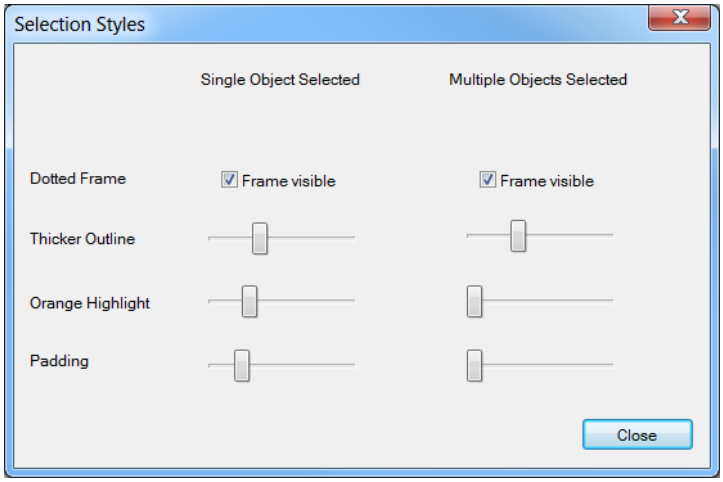
You can define for each page content type how the border, the print space or the different regions are

displayed. It is also possible to hide the outlines completely.

Visual Style of Object Selection

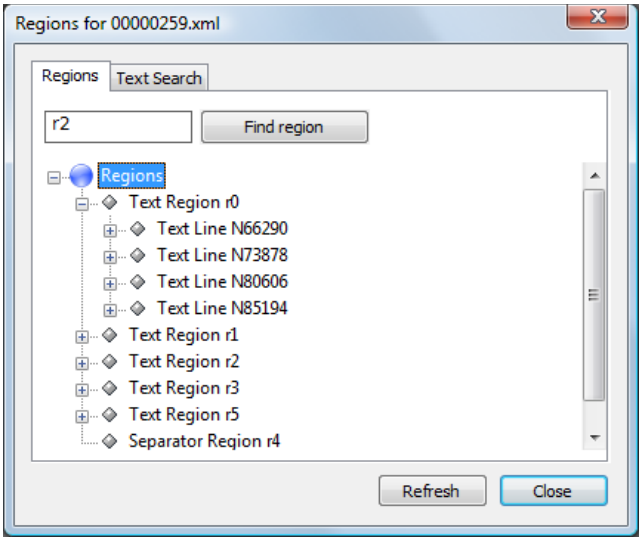
To change how selected objects are highlighted:

- Switch to the “View” toolbar tab
- Click on “Selection Styles” (opens the dialog shown below)
- Adjust the settings (changes take effect immediately)



Region Tree View

This view shows the hierarchy of page elements (regions, text lines, words and glyphs) in a tree. To open it, click ‘Regions’ in the toolbar panel called ‘Structure’ or press CTRL+F and select the ‘Regions’ tab. A dialog opens:



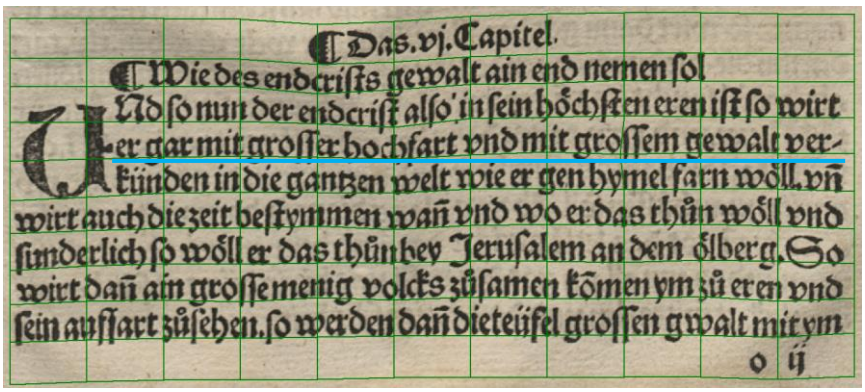
When selecting a page element within the tree, it will also be selected in the normal document view and vice versa.

It is possible to search for page elements (regions, text lines, words and glyphs) by their ID. To do so, enter the ID in the text field on the top and press ‘Find region’.

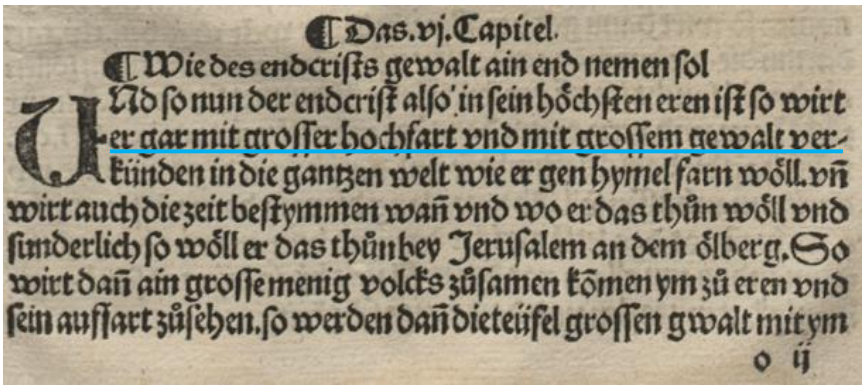
Dewarping

Dewarping is a grid-based method for geometric correction of document images. Aletheia allows creating and saving dewarping ground truth as well as load existing dewarping data from XML files.

Example of dewarping grid:



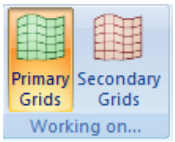
Dewarped image:



Several grids (non-overlapping) can be defined for one page. Aletheia also allows working on two independent sets of grids in parallel (e.g. for comparing ground truth against grid detection result).

To select the active set of grids:

- Use the buttons in the 'Working on…' toolbar panel



To select which set of grids is visible:

- Tick or untick the corresponding checkboxes in the 'View' toolbar panel

Loading and Saving Grids

Please note that the dewarping file format is not yet fully integrated into Aletheia. All file operations need to be carried out from the dewarping toolbar. It is not possible to load dewarping XML files from the Aletheia menu.

To load a set of grids from an XML file:

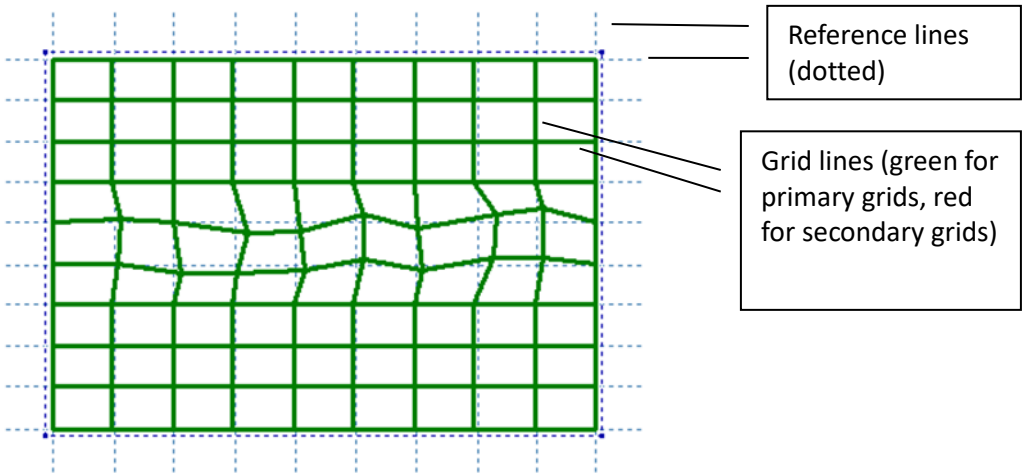
- Click on 'Open' in the 'File' toolbar panel
- Select the XML file

To save the currently active set of grids (primary or secondary):

- Click on 'Save'
- Select a file location

Working with Grids

A dewarping grid is a matrix of points that are connected to each other horizontally and vertically, forming grid lines.
In addition, reference lines can be defined. By default the reference lines are position at the average location of all corresponding grid points.

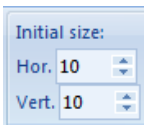


Note: Dewarping grids of the same type (primary/secondary) should not overlap.

Creating a grid

To create a new grid:

- Select the number of initial horizontal and vertical lines



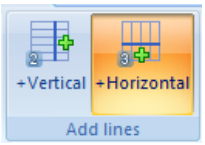
- Activate the ‘New grid’ tool from the ‘Grid’ toolbar panel

- Drag a rectangle where you want to position the grid

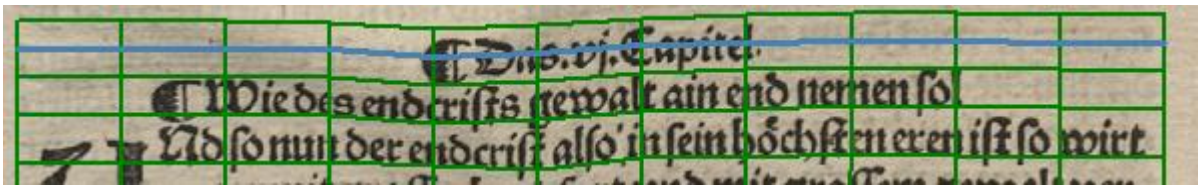
Modifying a grid

To add new lines:

- Activate one of the tools from the ‘Add lines’ toolbar panel (‘+Vertical’ or ‘+Horizontal’)



- Hover with the mouse cursor over the grid (a preview of the new line will be shown)

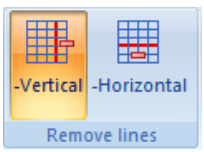


- Click left to add the line

Note: New lines will take the shape of the closest existing grid line.

To remove grid lines:

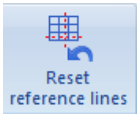
- Activate one of the tools from the ‘Remove lines’ toolbar panel (‘-Vertical’ or ‘-Horizontal’)



- Hover with the mouse cursor over the line you want to remove until it is highlighted
- Click left to remove the line

To reset reference lines:

- Select a grid
- Click ‘Reset reference lines’ in the ‘Remove lines’ toolbar panel



To move points or lines:

- Activate the ‘Edit’ tool from the ‘Basic tools’ toolbar panel
- Hover with the mouse cursor over a point, a line, or a reference line until it is highlighted
- Click left, hold, and drag the mouse to move the point or line
    - Note: By default the movement is confined to avoid overlapping grid lines. Press CTRL to disable the confinement.

“Lock”:

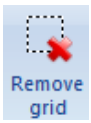
To avoid accidental horizontal movement of points, activate ‘Lock’. Then only vertical movement is possible.

Notes for reference lines:

- To move reference line hover the outmost part of the lines
- Once moved, reference lines change colour from blue to red to distinguish calculated lines from manually set lines

To delete a whole grid:

- Select the grid by clicking inside it (Hand tool or Select tool)
- Click on ‘Remove grid’ or press the delete key



Assistive grid creation

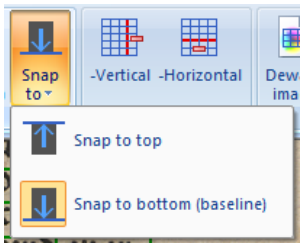
Grids can be defined more conveniently by using the assistive ‘Anchor point’ tool. To do so:

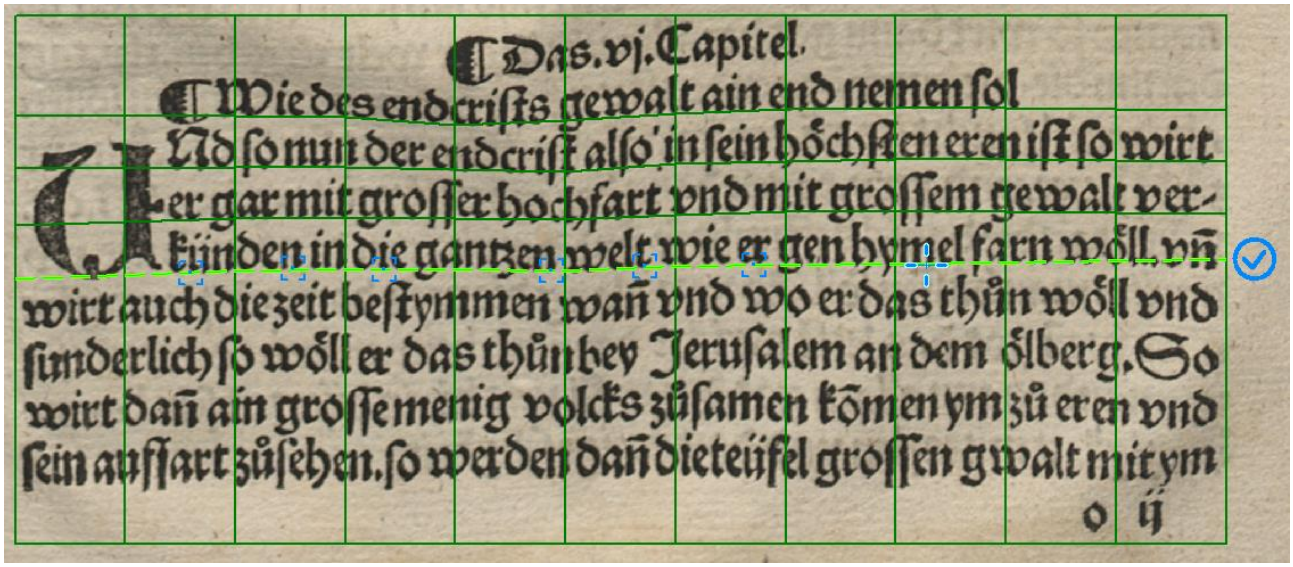
- Create an initial grid with only vertical lines using the ‘Initial grid’ tool from the ‘Assisted’ toolbar panel (tool identical to the ‘Add grid tool’)

- Activate the tool '+Horizontal (anchor points)'
- Add anchor points for one grid line by clicking left (a preview of the line is shown)
- Add the new line by clicking on the tick icon on the right side
- Repeat for the remaining grid lines

"Snapping":

Provided a black-and-white image is available, the tool 'snaps' to the bottom or top edge of foreground (e.g. text), when close to the edge. You can select in which direction to snap from the toolbar:

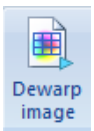




Dewarping an Image

If grids have been defined or loaded, the current document image can be dewarped:

- Select the image you want to dewarp (colour or black-and-white; Tab key)
- Optional: Tick 'Use reference lines' if you want to use the manually defined reference lines. Otherwise the average lines are used for dewarping.
- Click on 'Dewarp image' from the 'Run' toolbar panel



- Specify where to save the dewarped image

Note: If successful, the image should be opened after the dewarping is complete.

Validation

The validation can be used to make sure that a document layout has no major faults and that it meets the ground-truthing guidelines.

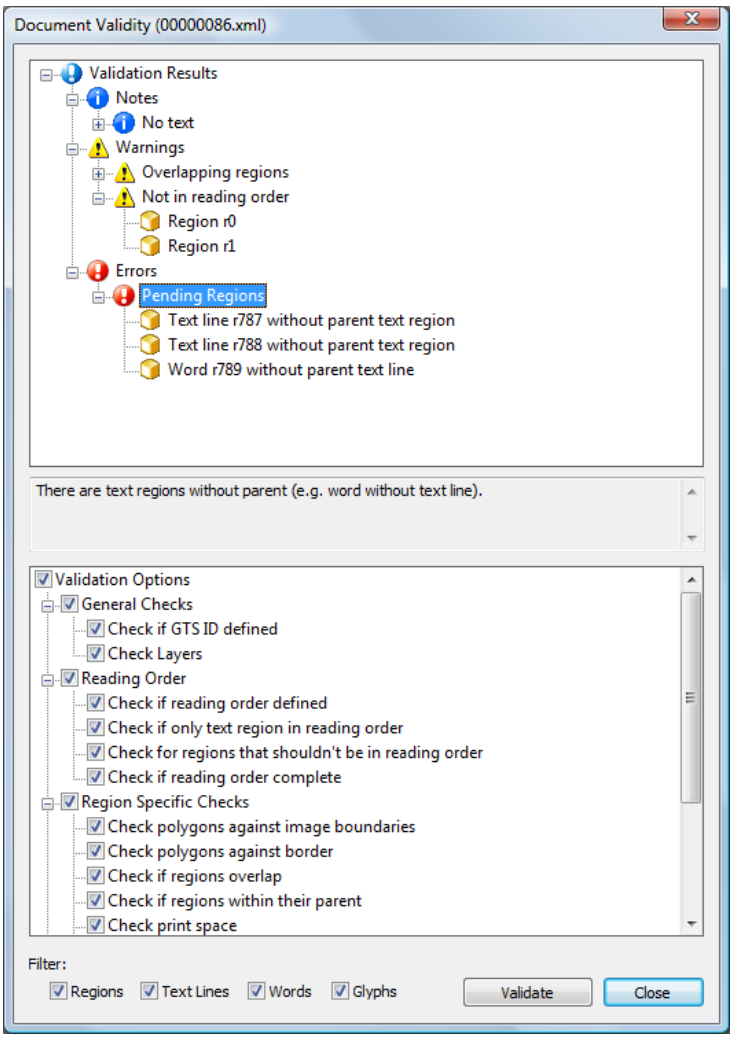
To validate a document layout:

- Switch to the Home category and click ‘Validation’ in the toolbar panel called ‘Quality’ (keyboard shortcut Ctrl+F1)

- Optional: Choose what to validate using the tree at the bottom
- Press ‘Validate’ to run the validation.

When finished, the results are presented as a tree (see the screenshot). For most nodes further details are displayed in the text filed beneath the tree, if they are selected. Clicking a node representing a region (block, text line, word, glyph), selects the region within the document image view. That way, problematic regions can be localized quickly.

To filter the tree to show only specific problems, select or deselect the validation tasks (at the bottom of the dialog) as desired and click ‘Validate’ again.
It is also possible to filter by region type (layout region, text line, word or glyph).



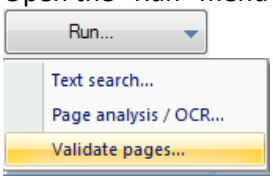
Check Box	**Message(s) if Invalid**	**Description**
Check if reading order defined	No reading order	No reading order is defined for the document.
Check reading order elements	Suspicious regions in reading order	There are one or more regions within the reading order that don't belong there (only paragraphs, headings, drop-capitals, catch-words and TOC-entries are supposed to be in the reading order).
Check if GTS ID defined	No GTS ID defined	The Ground Truth and Storage (GTS) ID is not defined in the document metadata.
Check layers	Layers incomplete	There are regions that are not assigned to a layer. (Only checked, if there is at least one layer.)
Check highlights	Highlight regions found	Highlight regions are special annotations to draw attention to ground truth problems. They should be removed eventually.
Check polygons against image boundaries	Polygon out of document boundaries	One or more points of a polygon are out of the document's boundaries.

Check polygons against border	Regions out of border	Some regions are partly outside the specified document border
Check if regions overlap	Overlapping regions	Some regions overlap each other
Check if regions within their parent	Regions not within parent	Some text regions are partly or completely outside their parent region
Check for intersecting polygon lines	Intersecting Polygon Lines	One ore more polygons have intersecting lines and therefore contain loops.
Plain text used instead of Unicode	No Unicode text defined	Some text regions have plain text defined but not Unicode text.
Check if text defined	No text	Some text regions have no text ground truth.
Find pending regions	Pending Regions	There are text regions without parent (e.g. word without text line).
Check if components completely inside region*	Components partly outside region	There are regions with connected components that are not completely inside the region polygon.
Print space related checks	Print Space	Checks if regions of type page-number, signature-mark, marginalia or catch-word are NOT within the print space
Check for deprecated characters	Deprecated Characters	Checks whether there are text elements with deprecated Unicode characters.
Check for replacement character	Replacement Character	Looks for the replacement character (Unicode +FFFD) in text elements.
Check for pending character	Pending Character	Looks for pending character(s) (Unicode +F51C) in text elements.
Check text content	Text Content	Checks for inconsistencies in the content of text elements (trailing line breaks etc.)

*Requires black-and-white image

To validate multiple pages at once:

- Open the “Run” menu in the page collection pane



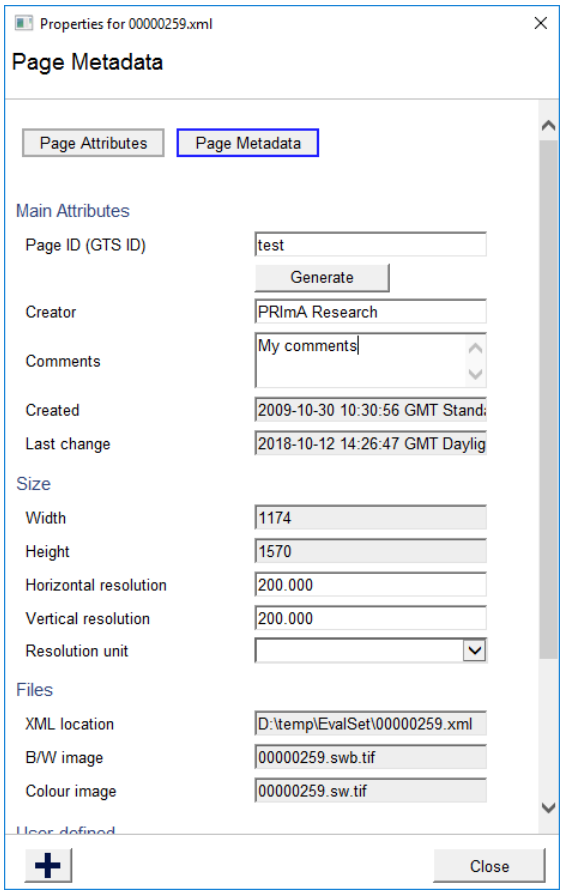
- Select “Validate pages”
- Choose what to validate
- Click on “Validate”

Metadata

Metadata is additional information that is not part of the document layout itself. It includes a ground-truthing ID, the creator of the document layout, comments and others. For user-defined attributes see the next subsection.

To view or change the metadata of a document:

- Switch to the Home category and click ‘Metadata’ in the toolbar panel called ‘Page’
- Edit the data (not all fields editable)
- Click ‘Close’



Fields:

Field	Description
GtsId	Ground-Truthing and Storage ID. This is an XML ID and it usually consists of one or more letters followed by a number. For detailed information see: http://www.w3.org/TR/2001/REC-xmlschema-2-20010502/
Creator	Name of the person who created the ground truth
Comments	Any additional notes concerning the document
Created	Creation date and time
LastChange	Date and time of the last modification
Width, Height	Dimensions of the document image

Colour image, Black-and-white image	File names of the document images

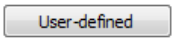
To get a proposal for the ID, click ‘Generate ID’. Aletheia then tries to construct an ID from the document file name.

Custom Metadata

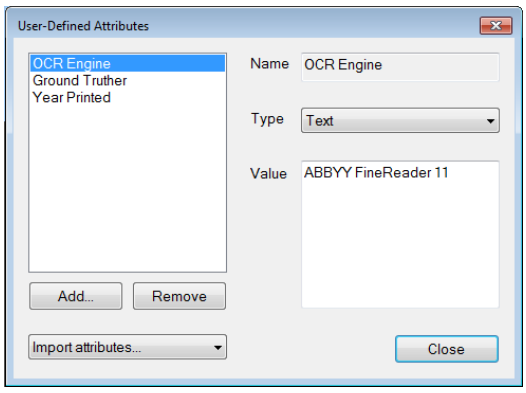
It is possible to add more metadata fields.

To add user-defined metadata attributes:

- Click on “User-Defined”

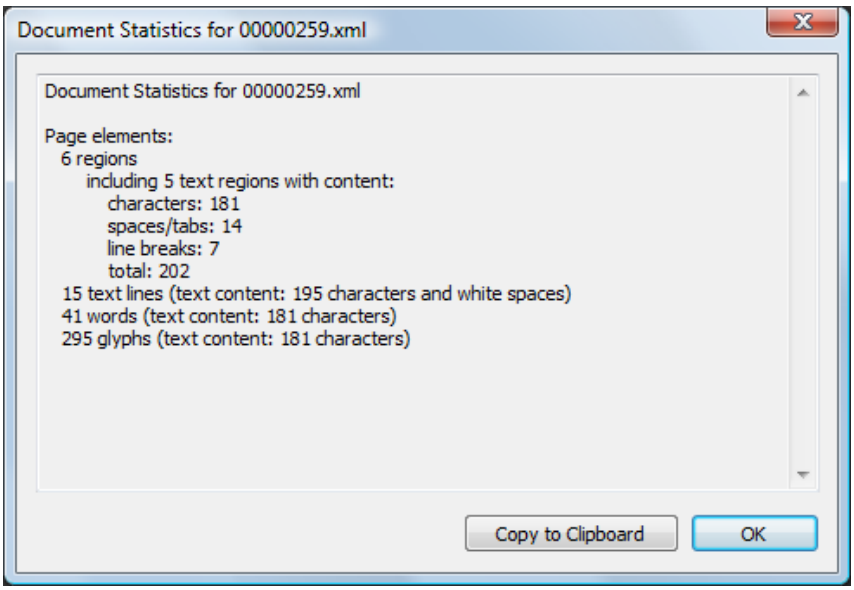
  

- In the dialog:
  - Add attributes:
    - Click on Add and enter a name
    - Select a type
      - Text (e.g. “Shakespeare”)
      - Integer number (e.g. “23”)
      - Decimal number (e.g. “8.5”)
      - Boolean (“true” or “false”)
  - Edit and attribute:
    - Select and attribute and select a new type or enter a new value
  - Remove an existing attribute:
    - Select an attribute and click on Remove
  - Save or load an attribute profile
    - Expand the “Import attributes” drop-down list
    - Select “Save…” to save your current attributes as profile for reuse
    - Select a saved profile to add all attributes of the profile



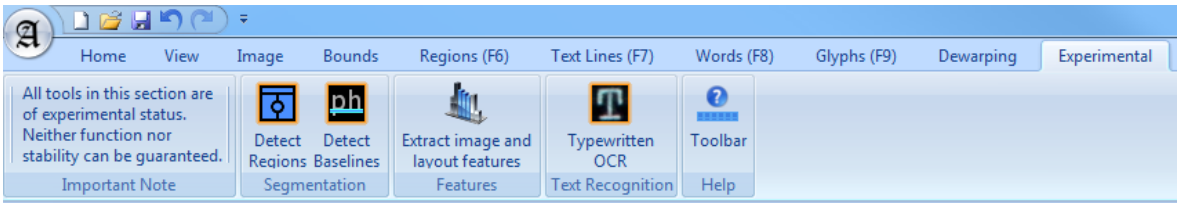
Statistics

To view some statistics on the current document switch to the Home category and click ‘Statistics’ in the toolbar panel called ‘Page’. A message box pops up:



Experimental Feature Section

The tools that are grouped under the “Experimental” toolbar category are not mature or optimised in any way. Function and stability of these features cannot be guaranteed. They are intended merely to showcase prototypes which may be made available – in a more optimised way – in context of research or other projects.

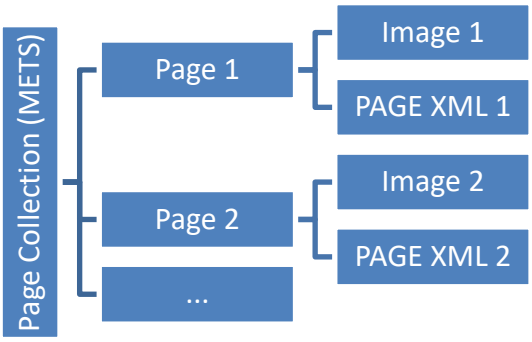


Page Collections

Since version 3.2, functionality for page collections is available, including:

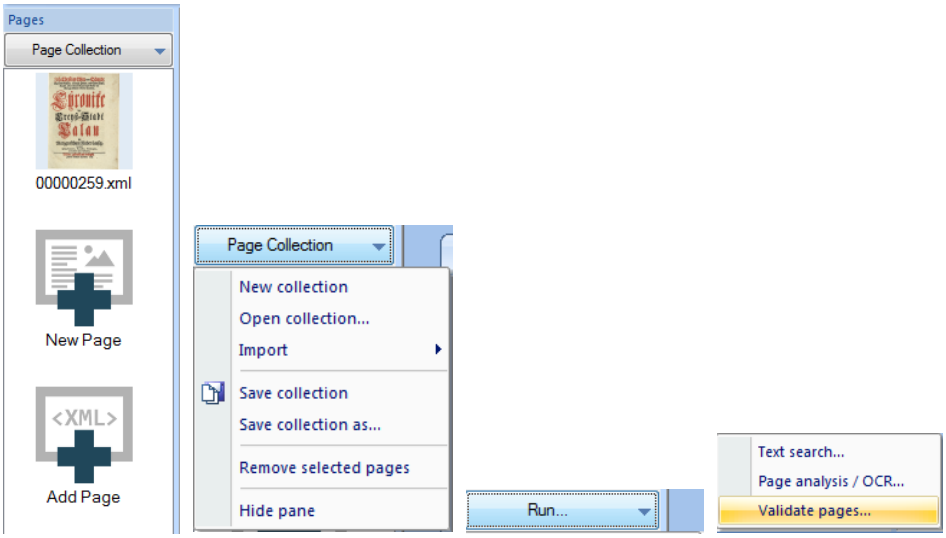
- Adding pages from images or existing page XML files
- Creating collections by importing images from a folder or multi-page XML files such as ALTO XML
- Saving / opening collections
- Running tools on all pages

A page collection can represent all pages of a printed document (e.g. a book) or a loose collection of unrelated pages. Page collections are stored as METS XML file, linking to the individual page images and content PAGE XML files:



In Aletheia, there is always exactly one active page collection. Each page that is opened or created new is automatically added to the current collection.

All page collection functions can be accessed via the dedicated tool pane:



If the page collection pane is hidden, it can be displayed from the Aletheia menu ("Show page collection pane"). The pane can be resized and moved to a different location.

The pane shows thumbnails of the open pages and placeholders for closed pages. Clicking on a page will open it and/or bring it to the foreground (if already open).

Creating, Opening, and Saving Page Collections

Note: Opening or creating a page collection will close the current page collection.

To **open** a saved page collection:

- Select “Open collection” from the page collection menu
- Choose a .metsx file (a METS XML file for Aletheia)
  OR
- Drag & drop a .metsx file from the Windows file explorer into Aletheia

To **save** the current page collection:

- Select “Save collection” or “Save collection as” (to save under a new name)
- (Select a file location)

To **import** pages from other collections:

- Open the “Import” sub-menu
- Choose an import source
- Choose the file or folder location

To **add** pages:

- Click on the “New page” to add a page from an image only or click on “Add page” to add an existing page content file
  OR
- Create or open a document in Aletheia as usual

Note: Pages that were not specifically added or imported, will be removed from the current collection if the page is closed, unless the page collection was saved in the meantime.

To **remove** pages from the current collection:

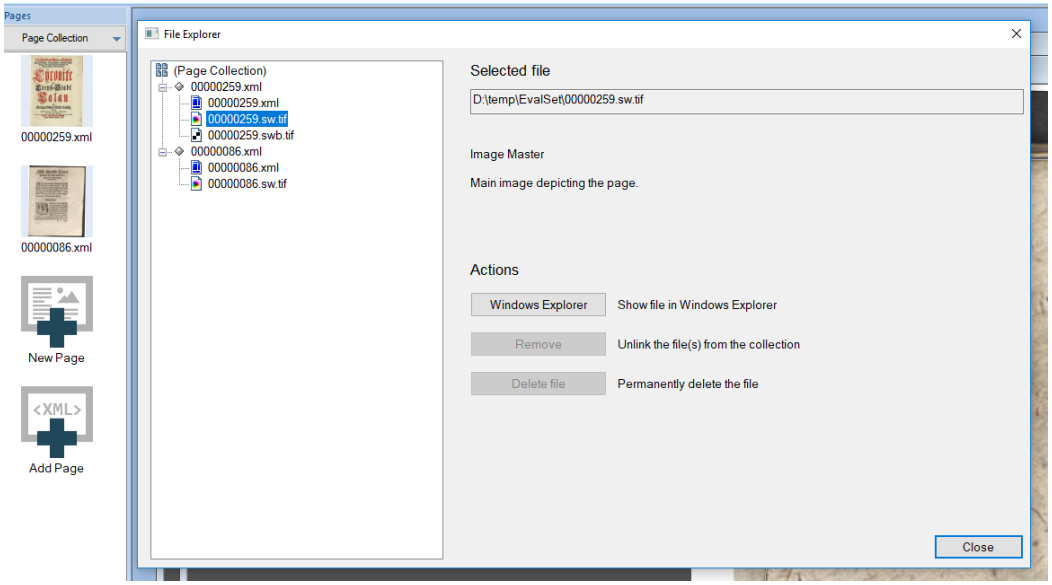
- Select one or multiple pages (using CTRL or SHIFT)
- Choose “Remove selected pages”
- (The removed pages will be closed)

Note: Currently, there is no undo/redo for page collection operations.

File Explorer

To view all XML and image files of the current pages:

- Open the Page Collection menu
- Click on File Explorer
- The dialog opens:

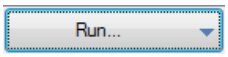


Detailed information and actions are shown when selecting an item in the tree on the left.

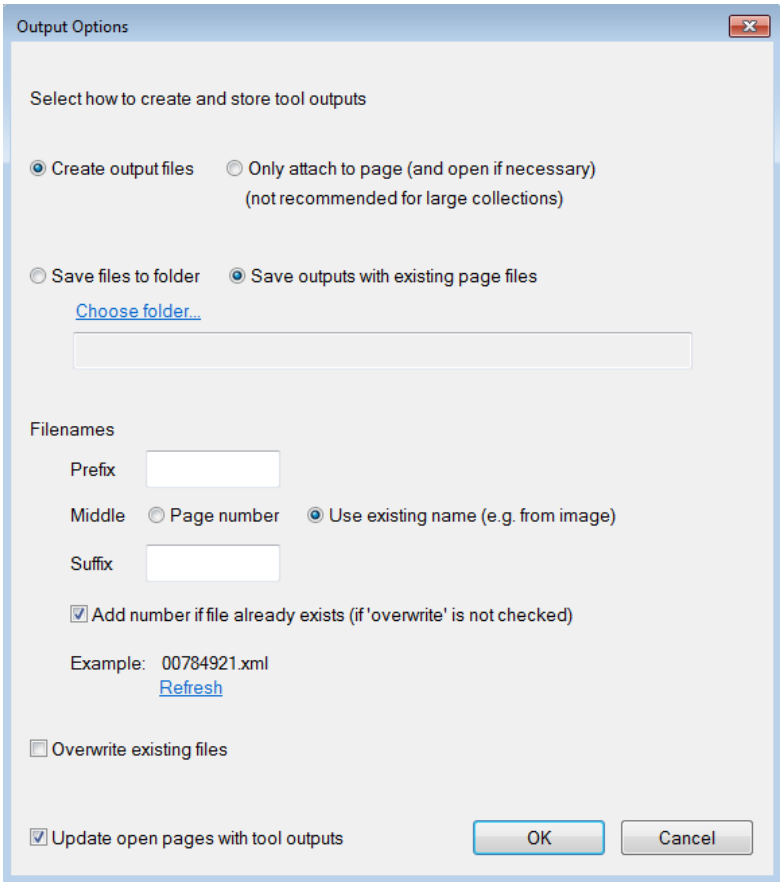
Running Tools for All Pages

To run multi-page operations:

- Select the “For all pages” option in the corresponding tool dialog
  OR
- Click on “Run…” at the bottom of the page collection pane and select the operation



Some operations require settings for what to do with tool outputs (such as an OCR result):



Available options:

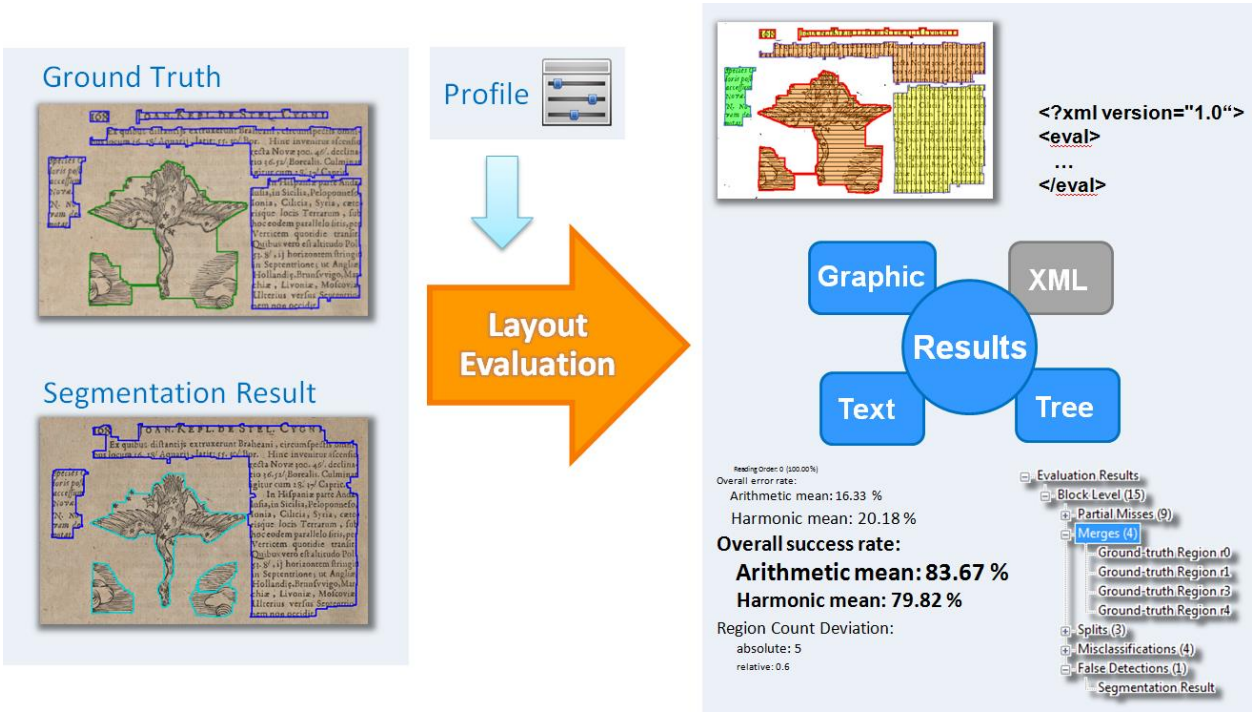
- "Create output files": Create and store files for the tool outputs
- "Only attach to page": Don't create files, only open the results within Aletheia (attached to each page). This will open all pages of the collection, if not open yet.
- "Save to folder": Store the created files in a specific folder (click on choose folder to specify the location)
- "Save with existing page files": Determine the output location from other page files. When performing OCR, this will be the image file location.
- Filename pattern: Defines how output files will be named. The general pattern is [prefix]<middle>[suffix] [number].extension. Use "Refresh" to see an example of your current settings.
- "Overwrite": Choose this option to replace existing files with the tool outputs. Make sure to have a backup of the existing files if you are not sure.
- "Update open pages": Select this option to replace the page content with tool outputs if the corresponding page is open in Aletheia. This is equal to running the tool for the current page on its own.

Performance Evaluation

Aletheia now contains some functionality from the [PRImA Layout Evaluation Tool](). In addition, evaluation for the long-standing ICDAR Layout Analysis and Recognition competitions organised by PRImA was included.

Layout Evaluation

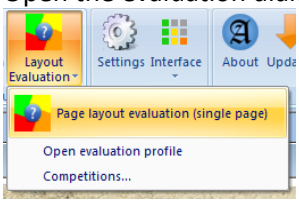
Layout evaluation is used to benchmark results of layout segmentation methods and uncover specific problems of the algorithms to help developers to improve them. As input the ground truth XML file, the segmentation result XML file and the black-and-white document image are required. The colour image 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:

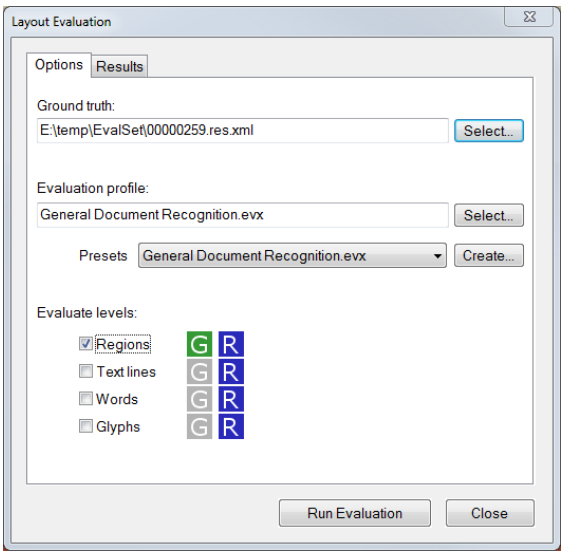


In-depth Evaluation for a Single Page

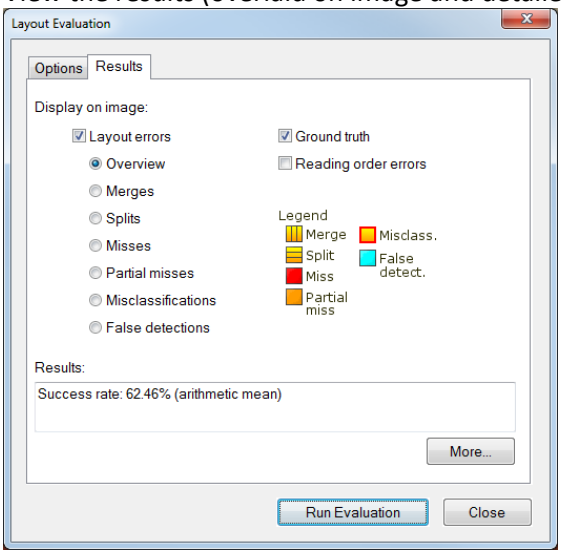
To compare your page segmentation results against ground truth:

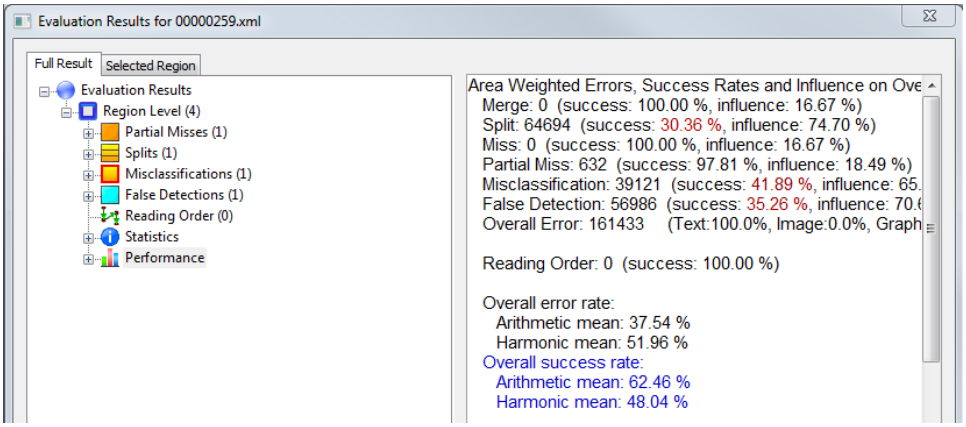
- Open the segmentation results in Aletheia
- Open the evaluation dialog (Toolbar Home - Quality)





- Select the ground truth file
- Choose an evaluation profile from the drop-down list, select a profile file, or create a new profile (save it and select the file)
- Select what you want to evaluate (regions/zones, text line objects, word objects, and/or glyph objects) (the coloured icons next to the check boxes indicate what data is available in your files: G = Ground Truth, R = Segmentation Result)
- Click on “Run Evaluation”
- View the results (overlaid on image and detailed in results dialog with tree view):

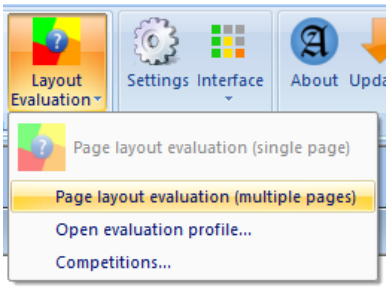




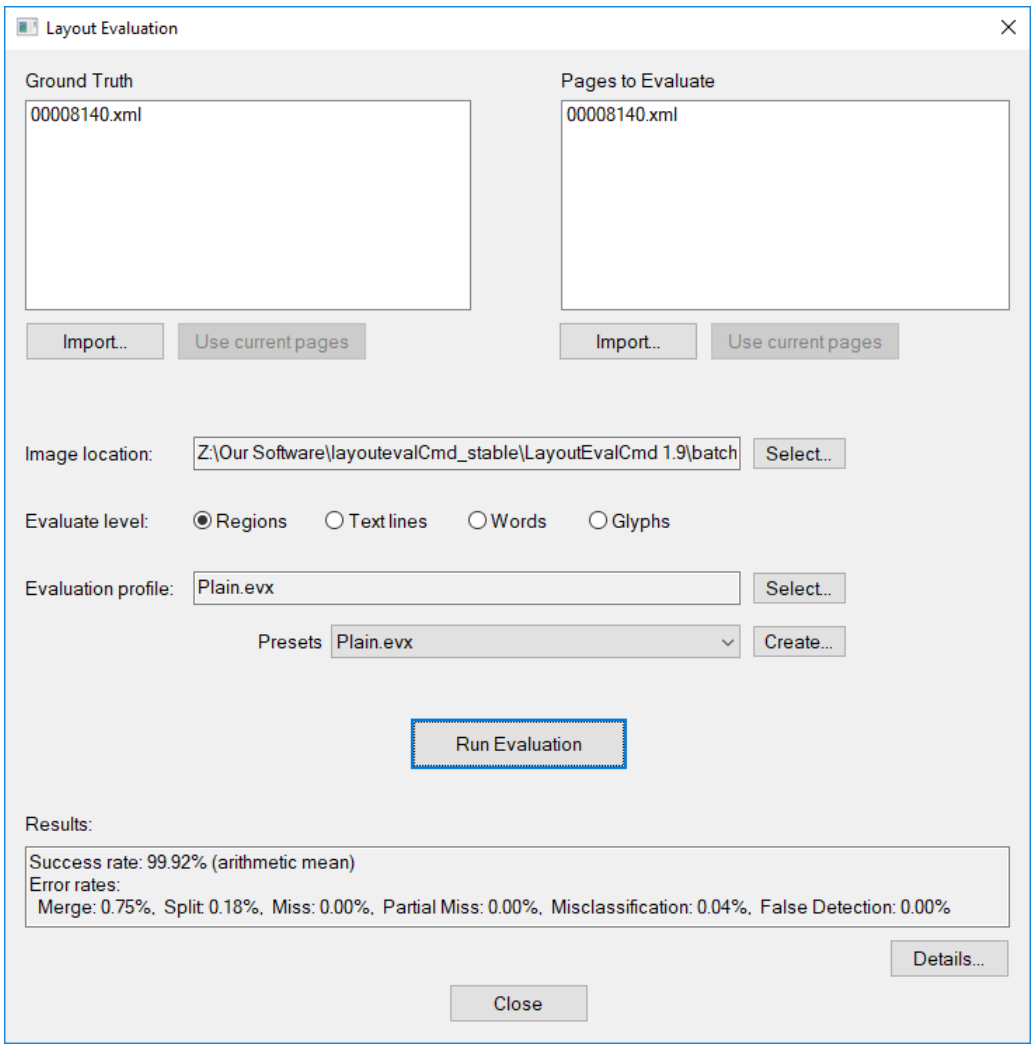
Batch Evaluation of Multiple Pages

To evaluate multiple pages in one go:

- Select “Page layout evaluation (multiple pages)” from the toolbar menu



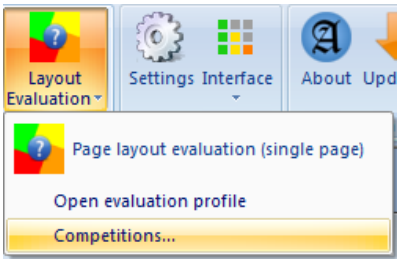
- Import ground truth page files and page files to evaluate
    - Select a folder to import all XML files from within the folder
    - Choose “Use current pages” to import all pages from the current page collection in Aletheia
- Optional: select a folder with images (if not specified, the XML folders will be searched for the document images)
    - Ideally provide bitonal black-and-white images
    - Colour or greyscale images will be converted to black-and-white internally
- Select one of the provided evaluation profiles or open a custom one
- Click on Run Evaluation
- (Export comma-separated values with detailed evaluation results using the “Details...” button)

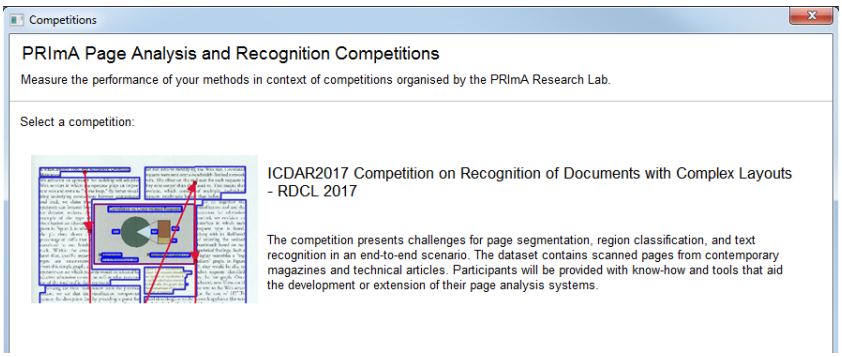


Competitions

To evaluate your methods in context of competitions:

- Open the competitions dialog (Toolbar Home - Quality)
- Follow the instructions




Competitions
PRImA Page Analysis and Recognition Competitions
Measure the performance of your methods in context of competitions organised by the PRImA Research Lab.
Select a competition:
ICDAR2017 Competition on Recognition of Documents with Complex Layouts - RDCL 2017
The competition presents challenges for page segmentation, region classification, and text recognition in an end-to-end scenario. The dataset contains scanned pages from contemporary magazines and technical articles. Participants will be provided with know-how and tools that aid the development or extension of their page analysis systems.

Customisation and Settings

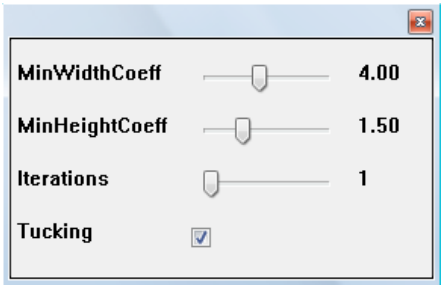
The user can customise several settings of Aletheia. Some of them are saved automatically (e.g. window positions) and others have to be adjusted manually.

Window Positions and Sizes

Aletheia automatically saves position and size of the main window and all dialogs.
To reset the positions switch to the View toolbar tab and click ‘Reset Dialogs’ in the toolbar panel called ‘Window’.

Algorithm Parameters

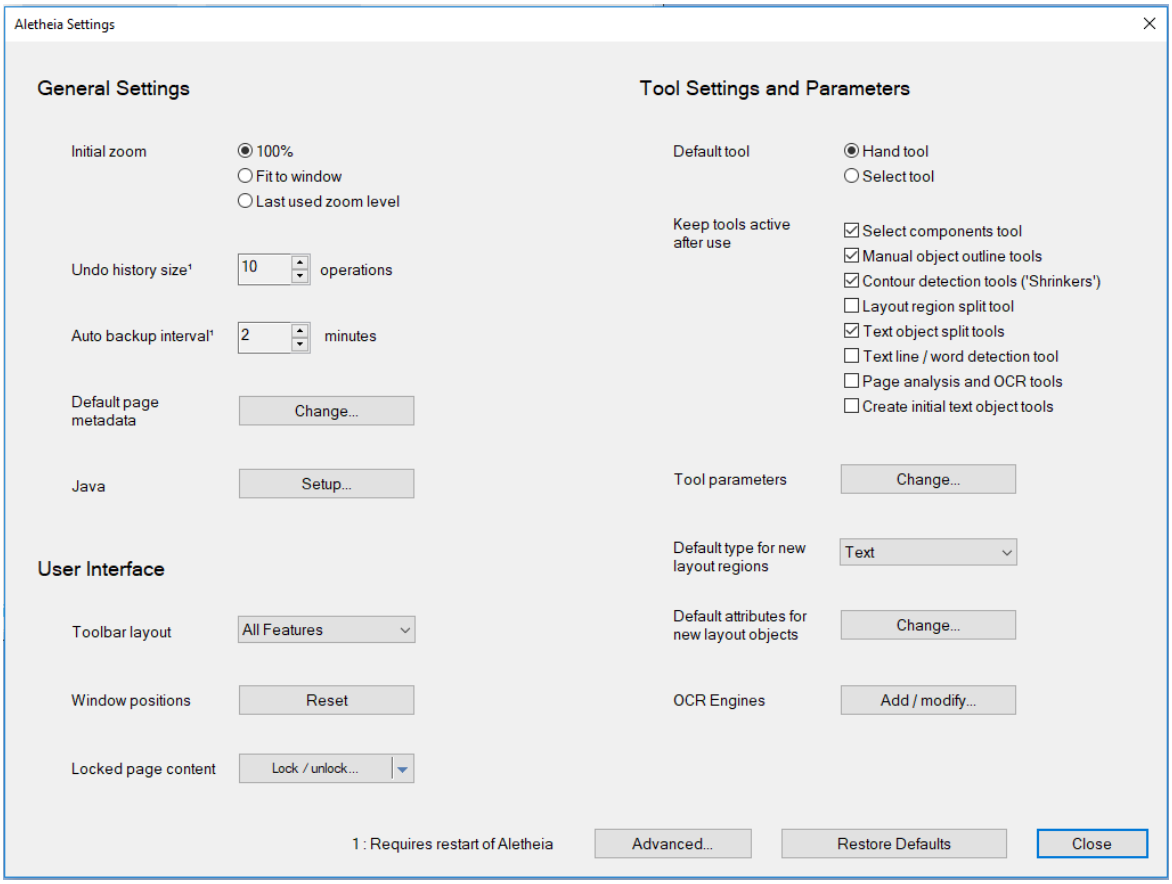
The parameter setting for algorithms such as for the ‘Text Region Tool’ is also saved automatically into the user settings. The next time the tool is used, the parameters will have the same value as the last time.
The lower and upper limits of number parameters (displayed as sliders) can be changed within the settings dialog (see next section). Though, only experienced users should change the boundaries.



The Settings Dialog

To adjust the settings of Aletheia switch to the Home category and click ‘Settings’ in the toolbar panel called ‘Customisation’.

The settings dialog opens:



- Click on “Advanced…” to see all available settings

Setting	**Description**
Auto backup interval	Defines the interval in minutes to automatically save all documents to a temporary folder in the background. The saved files can be used to recover documents after a system failure.
Undo history size	Number of recorded operations for undo and redo.
Initial zoom / default zoom	Specifies the zoom level to be used when opening a document. There are three options: 100%, ‘Fit to window’ and ‘Last used zoom level’
Default tool	Specifies the tool to be used when opening a document. There are two options: ‘Hand tool’ and ‘Select tool’
Refine outlines	Specifies if to try to refine region outlines when loading non-PAGE files such as ALTO. If set to ‘false’, the original outlines are kept.
Convert to isothetic on save	(disabled)
Plain text editable	Defines if the plain text input field is enabled for text regions. It is recommended to use the Unicode text field.
Min component area	Specifies a minimum area for connected components. Smaller components will be ignored. Connected components are used for several tools in Aletheia. Choosing a higher value speeds up loading images and lessens the memory consumption. However, if the value is chosen too big, the ground-truthing process may be complicated.
Mouse drag start	Mouse movement (in pixel) to determine if the mouse is dragged by the user. Only increase this if mouse dragging is often wrongly triggered (instead of a mouse click).
Poly point dist	Minimum distance between two polygon points (in pixels) when drawing with the mouse. Smaller values lead to more detailed polygons.

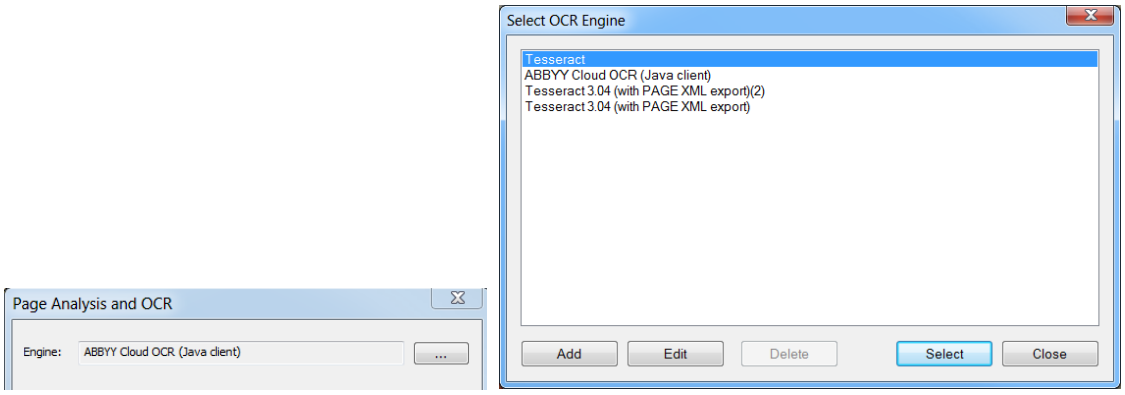
Poly finish by double-click	Enable or disable to finish drawing a polygon by double-click.
Default page metadata / Metadata default values	Values that will be automatically written to the meta data when creating a new document.
Tool / algorithm parameters	Default parameters for several tools of Aletheia. **Note:** Though it is possible to change the lower and upper limit of number parameters, it is not recommended. Changing the limits without knowledge of the internal structure of the algorithm can cause program failures when running the algorithm.
Keep tool active / Stickiness of tools	Options to either keep specific tools active after running them or to switch to the default tool. For example, if the option for manual region tools is set to TRUE (sticky), the rectangle tool will still be active after creating a rectangular region. Another region can then be created immediately without having to activate the tool again.
Default region type and default layout object attributes	Values that will be used when creating a new region or other layout object.
Toolbar layout	Changes the layout of the main toolbar. This setting is the same as the "Interface" drop-down menu in the toolbar.
Window positions	Resets position and size of all dialogs.
Locked page content	Allows to lock/unlock certain aspects of the page content. Locked content cannot be edited.
OCR Engines	Change the way Aletheia runs page layout analysis / OCR. Custom engines can be linked via command line interface.
Java	Check if Java is available on the current machine, and, if not, link to the local Java installation. Java is required for some features.

To restore the original settings, press the 'Restore Defaults' button.

OCR Engines

The OCR engine Aletheia uses to perform page analysis and text recognition is interchangeable. The default is Tesseract OCR version 4.0 (see https://github.com/tesseract-ocr). Tesseract 3.04 is also bundled with Aletheia.

OCR engines can be modified, added or deleted via the settings dialog or from within the page analysis / OCR dialog:



To change the OCR engine:

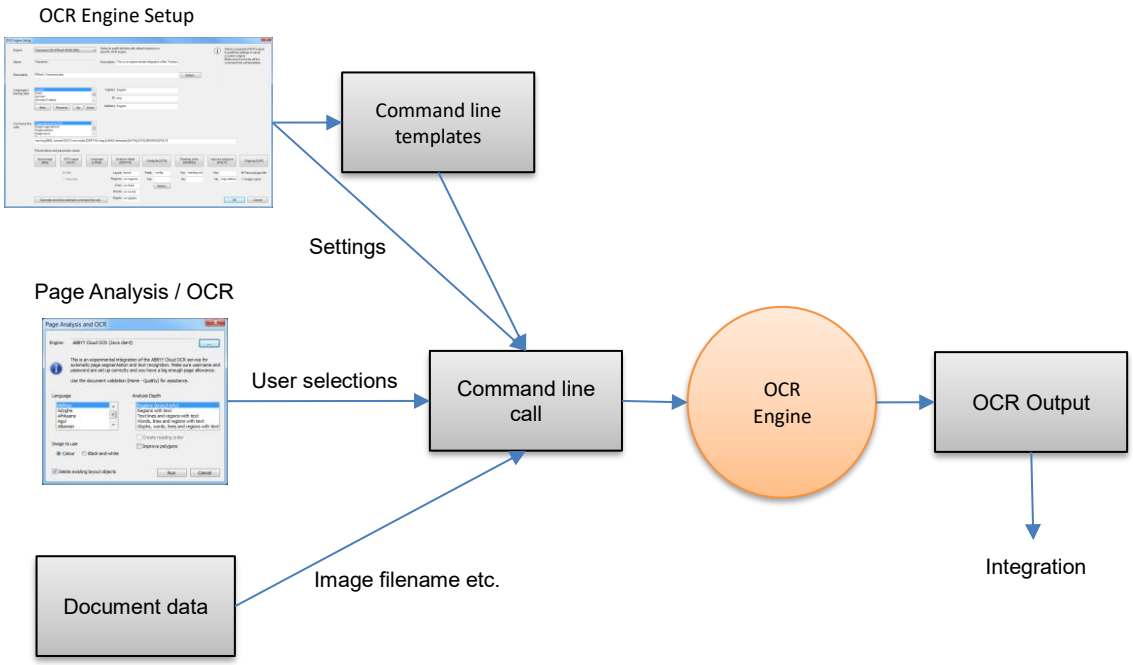
- Open the "Select OCR Engine" dialog ("…" button in Page Analysis / OCR dialog or "OCR Engines"

section in settings dialog)

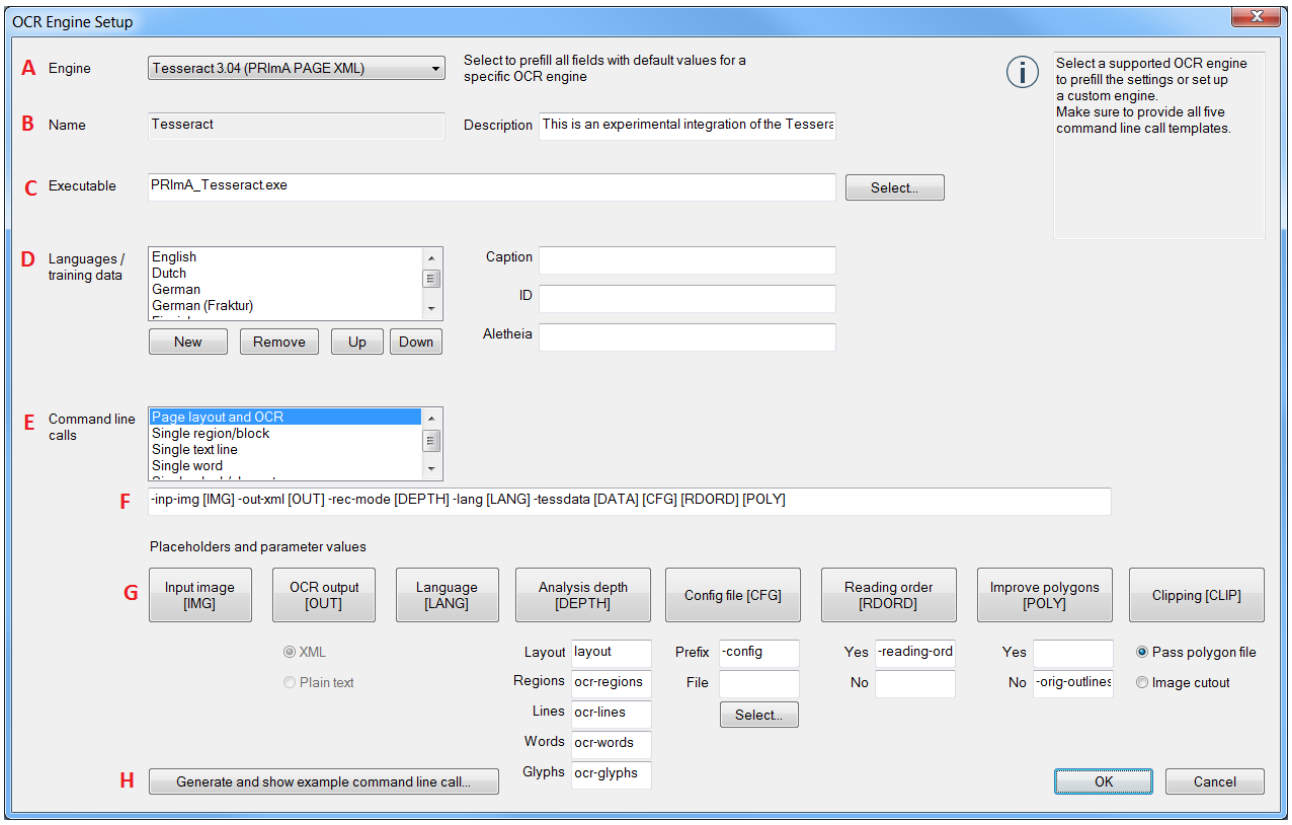
- Select an OCR engine from the list of available engines
- Click on “Select”

Adding an OCR Engine

OCR engines are linked via command line interface. An engine setup contains command line call templates and various settings. When performing page analysis / OCR, Aletheia fills in a template with data from the Page Analysis / OCR dialog and information about the current document. The so created command line call is used to run the OCR engine. The OCR results are loaded, interpreted and integrated into the current page the user is working on.



The OCR Engine Setup Dialog

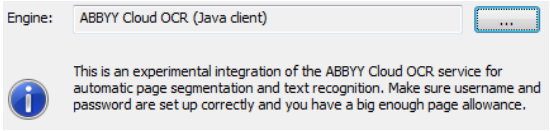


A: Engine selection

- Select a pre-configured engine to fill in settings and languages

B: Name and description

- Used as display name and filename for the engine settings (don't use special characters that are not allowed in filenames!)
- Name and description will also be shown in the Page Analysis / OCR dialog

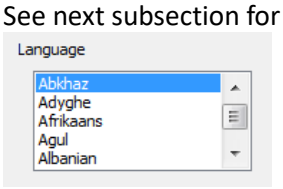


C: Executable

- OCR engine executable that is to be called
- If no path is specified, the Aletheia installation root will be used
- This can also be a Java call, for instance (e.g. java -jar c:\temp\abbyyCloudOCR.jar)

D: Languages

- Languages that are available in the Page Analysis / OCR dialog
- See next subsection for details



- A separator can be specified which is used if multiple languages are selected by the user when running page analysis / OCR (for Tesseract this is '+', for example)

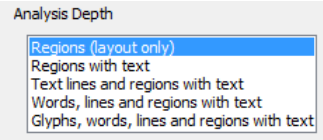
E, F: Command line templates

- Five templates for analysis / OCR on page level, region level, text line level, word level and glyph level

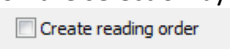
- Select an entry in the list box (E) to view or modify the command line call template (F)
- Use placeholder (G) which will be filled in dynamically from Aletheia when running OCR

G: Placeholders

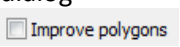
- Convenience buttons to insert placeholders into the current command line template
- [IMG] = Image file to be passed as input for the OCR engine
- [PAGEINDEX] = Index of page if working with multi-page images
- [OUT] = OCR result output file
- [LANG] = Language input argument for the OCR engine ("ID" field of the user-selected language)
- [DEPTH] = User-selected analysis depth, uses on of the five commands below the button



- [CFG] = Config file, will be replaced with '<prefix> "<file>" ' if a file is specified. For Tesseract, a permanent config.txt file can be placed in the Tesseract folder within Aletheia. This file will be used except if overridden by an explicit file in the field below the [CFG] button
- [RDORD] = Reading order command. Uses one of the two commands below the button, depending on the selection by the user in the Page Analysis / OCR dialog



- [POLY] = Command to improve polygons or keep the original OCR output. Uses one of the two commands below the button, depending on the selection by the user in the Page Analysis / OCR dialog



- [CLIP] = Clipping polygon file. Temporary file with polygon points that is passed to the OCR engine to restrict OCR to the inside area of the polygon. The file format is one point per line with <x>,<y>. If "Image cutout" is selected, a snippet of current document image will be sent to the OCR engine as [IMG].
- [DATA] = Tesseract-specific placeholder for tessdata folder. Uses the parent folder of the specified executable, appended with "\tessdata"

H:

- Click to view and example command line call for the current template

Adding Languages for Tesseract Page Analysis and OCR

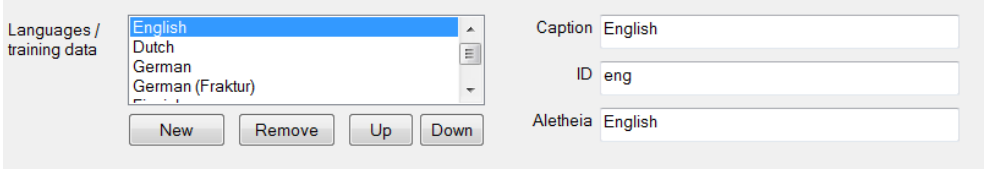
The data files for supported languages are located in the Aletheia main folder under:
"bin\data\tesseract\tessdata" for Tesseract 3.04 and
"bin\data\tesseract4\tessdata" for Tesseract 4.0.

Data files for additional languages are available on the Tesseract GitHub repository. Note that Tesseract 4 uses data files for scripts (in addition to the language data files .traineddata).

To add or delete languages:

- Open the "Select OCR Engine" dialog ("…" button in Page Analysis / OCR dialog or "OCR Engines" section in settings dialog)
- (Select Tesseract)
- Click on "Edit"
- Click "New" to add a language (will be initialised with placeholder name)
- Select a language and change the fields on the right to modify it:
    - "Caption" is the name that is displayed in the Page Analysis / OCR dialog

- “ID” is the Tesseract ID that used for the data file and command line argument
- “Aletheia” is the PAGE XML name of the language that is used to preselect the language in the Page Analysis / OCR dialog if OCRing a text region (if the user specified a language before)
-



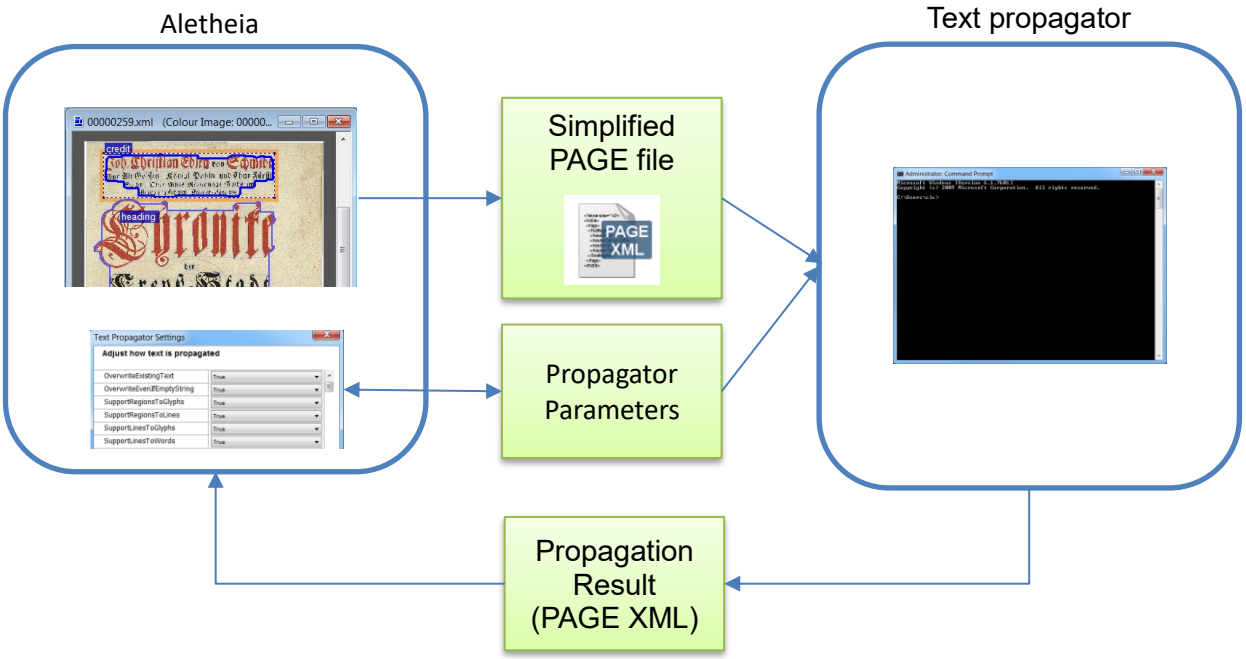
- Use the “Up” / “Down” buttons to change the sort order of the languages

To manually add new languages for the integrated Tesseract page analysis and OCR:

- Put the language data files in <Aletheia>\bin\data\tesseract\tessdata
- Modify the settings file <Aletheia>\bin\data\tesseract\languages.xml:
  - Add one <parameter> element for each language using following attributes:
    - ‘type’ = 4
    - ‘sortIndex’ = <intended position within the dialog list box>
    - ‘name’ = <display name>
    - ‘value’ = <Tesseract language ID (e.g. ‘eng’)>
    - <Description> (Text element) = <Aletheia language ID (e.g. ‘English’)> (optional)

External Text Propagation

Aletheia’s default text propagation method can be replaced by an external tool. Communication between Aletheia and the external tool is done via command line interface and input/output files (see figure).



Command line arguments:

[Propagator executable] <PAGE XML in> <PAGE XML out> <Parameter file> <Propagation mode>

- Propagator executable: Can be any kind of command line tool (e.g. Java)
- PAGE XML in: This is the full path to a temporary PAGE XML file that Aletheia creates when the user propagates text. Only the relevant objects (regions/lines/words/glyphs) are copied.
- PAGE XML out: This is the file path to where the external propagator is expected to save the propagation result.
- Parameter file: XML file containing parameters for the text propagator.
- Propagation mode: Is used to specify from which text object level to propagate to which other level. One of:
    - RegionsToGlyphs
    - RegionsToLines
    - LinesToRegions
    - LinesToGlyphs
    - LinesToWords
    - WordsToRegions
    - WordsToLines
    - WordsToGlyphs
    - GlyphsToRegions
    - GlyphsToWords

Error handling:
In case of a propagation error, the external tool should return an error code > 0 (e.g. “System.exit(2)” in Java). Aletheia will then display the error code and any output from stdout and stderr in the propagation dialog.

Simplified PAGE XML input and output:
Aletheia creates a copy of the original page containing only the relevant text objects. Which objects are copied depends on the propagation operation (mode). The external propagator should load the XML file, perform the propagation, and save the result using the specified output file path. Object IDs should not be changed, otherwise Aletheia will not be able to reincorporate the propagation results.

Parameter file:
The parameter XML file contains propagator settings that can be changed from within Aletheia. A parameter file will be created automatically, when a new propagator is set being set up, but a propagator can also be added by importing an existing parameter file. While additional, propagator-specific parameters can be added, two parameters should be supported by all external propagators:

- OverwriteExistingText (ID 101): Boolean parameter to indicate whether the user want’s existing text content to stay untouched (not to be overwritten by propagation)
- OverwriteEvenIfEmptyString (ID 102): Boolean parameter to indicate that, if “overwrite existing text” is set to true, the propagator should overwrite existing text even if the text that is propagated is empty

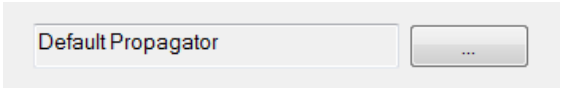
For external propagators, the parameters should also contain a text parameter called “Executable” (ID 11). This parameter should contain the full call for the propagator executable (including quotation marks for paths if required).

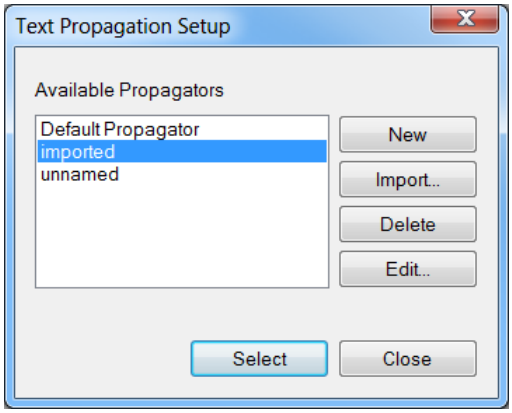
Example parameter file:

```
<?xml version="1.0" encoding="utf-8"?>
<Parameters>
	<Parameter	type="4" sortIndex="11" version="0" name="Executable"
			isSet="false" visible="true" value="…" id="11" textType="0"/>
	<Parameter	type="2" sortIndex="101" version="0" name="OverwriteExistingText"
			readOnly="false" isSet="true" visible="true" value="true" id="101"/>
	</Parameter>
	…
```

```
<Parameters>
```

Setup:
Propagators can be set up from within the text propagation dialog, using the "…" button. The "Text Propagation Setup" dialog allows to create, import, edit, delete and select a propagator. Aletheia comes preconfigured with an internal default text propagator. Make sure when adding a new propagator to avoid special characters for the name since it is used as filename to store the propagator's parameters.





Command Line Interface

It is possible to load a specified XML file together with an image via a command line call. The format is the following:

```
<xml file path>|<image file path>[|<object ID>]
```

Example:

```
Aletheia.exe "E:\temp\0001.xml|E:\temp\0001.jpg|N66290"
```

If an object ID is provided, Aletheia will highlight the corresponding page content object after the file has been loaded.

The image file path can be an incomplete path to a specific folder. If so, the image path is determined by combining the partial path with the image filename stored in the XML file.

Example:

```
Aletheia.exe "E:\temp\0001.xml|F:\images"
```

Keyboard Shortcuts

General

Shortcut	Description	Scope
Ctrl + N	New document	
Ctrl + O	Open document	
Ctrl + Q	Quick Open	At least one active document
F3	Go to document image	
F4	Go to document border	
F5	Go to print space	
F6	Go to regions	
F7	Go to text lines	
F8	Go to words	
F9	Go to glyphs	
+ (num)	Zoom in	
- (num)	Zoom out	
* (num)	Zoom to 100%	
/ (num)	Zoom to fit document into window	
. (num)	Zoom so that the width of the page fits the window.	
Cursor keys	Scroll the document	
Escape	Switch to default tool (hand or select)	
Ctrl + Z	Undo	
Ctrl + Y	Redo	
Ctrl + F1	Open validation dialog	
TAB	Toggle black-and-white / colour image	
Space	Switch to hand tool or switch back to the previous tool	
RETURN / ENTER	Finish polygon	When drawing polygon point by point
CTRL + -	Insert soft hyphen	Text dialog
H	Highlighter tool	

Image View

Shortcut	Description	Scope
Delete	Removes selected components from the black-and-white image	

Bounds View

Shortcut	Description	Scope
F4	View and edit the border	
F5	View and edit the print space	
P	Create polygonal border or print space	
R	Create rectangular border or print space	
Delete	Delete border or print space	

Region View

Shortcut	Description	Scope
F1	Switch to 'Select' tool	
F2	Switch to 'Edit' tool	
F10	Open dialog for region properties	
F11	Open text input dialog	
F12	Open reading order dialog	
Ctrl + A	Select all regions	
Ctrl + C	Copy text content to Windows clipboard; Copy region(s)	Region(s) selected
Ctrl + V	Paste previously copied region(s) (activates tool with preview)	
Delete	Deletes the selected region(s)	Select or hand tool
	Deletes multiple polygon points	Edit tool + points selected
	Deletes a polygon point	Edit tool + mouse hover over polygon point
Page Down	Apply region properties and select next region	Properties or text input dialog open
	Select next region	Neither properties nor text input dialog open
Page Up	Apply region properties and select previous region	Properties or text input dialog open
	Select previous region	Neither properties nor text input dialog open
Ctrl + ⮌⮊⮉⮋	Select neighbour region	
Ctrl + F	Open search dialog	
Ctrl + T	Enable/disable text overlay	
1	Fine contour rectangle tool	
2	Fine contour polygon tool	

3	Coarse contour rectangle tool	
4	Coarse contour polygon tool	
S	Select components tool	
C	Create region from components	Components selected
R	Create rectangular region	
P	Create polygonal region	
I	Create isothetic polygonal region	
A	Adjust region outline to text lines	
N	Split region tool	
M	Merge selected region	
O	OCR selected regions	Regions selected
Ctrl + B	Converts the outlines of selected regions to their bounding box	
Ctrl + I	Converts the outlines of selected regions to isothetic format	

Text Line View

Shortcut	**Description**	**Scope**
F1	Switch to 'Select' tool	
F2	Switch to 'Edit' tool	
F10	Open dialog for text line properties	
F11	Open text input dialog	
Ctrl + A	Select all text lines	
Ctrl + C	Copy text content to Windows clipboard; Copy text line(s)	Text line(s) selected
Ctrl + V	Paste previously copied text line(s) (activates tool with preview)	
Delete	Deletes the selected line(s)	Select or hand tool
	Deletes multiple polygon points	Edit tool + points selected
	Deletes a polygon point	Edit tool + mouse hover over polygon point
Page Down	Apply text line properties and select next line	Properties or text input dialog open
	Select next text line	Neither properties nor text input dialog open
Page Up	Apply text line properties and select previous line	Properties or text input dialog open
	Select previous text line	Neither properties nor text input dialog open

Ctrl + ⮜⮞⮝⮟	Select neighbour text line	
Ctrl + F	Open search dialog	
Ctrl + T	Enable/disable text overlay	
1	Create initial text line	
2	Split tool	
3	Split (cut) tool	
4	Contour detection (start from rectangle)	
5	Contour detection (start from polygon)	
S	Select components tool	
C	Create text line from components	Components selected
R	Create rectangular text line	
P	Create polygonal text line	
I	Create isothetic polygonal text line	
M	Merge selected text lines	
O	OCR selected text lines	Lines selected
A	Adjust text line outline to words	
Ctrl + B	Converts the outlines of selected text lines to their bounding box	
Ctrl + I	Converts the outlines of selected text lines to isothetic format	
D	Auto detection tool	
B	Create text region from text lines (bottom-up)	

Word View

Shortcut	**Description**	**Scope**
F1	Switch to ‘Select’ tool	
F2	Switch to ‘Edit’ tool	
F10	Open dialog for word properties	
F11	Open text input dialog	
Ctrl + A	Select all words	
Ctrl + C	Copy text content to Windows clipboard; Copy word(s)	Word(s) selected
Ctrl + V	Paste previously copied word(s) (activates tool with preview)	
Delete	Deletes the selected word(s)	Select or hand tool
	Deletes multiple polygon points	Edit tool + points selected

	Deletes a polygon point	Edit tool + mouse hover over polygon point
Page Down	Apply word properties and select next word	Properties or text input dialog open
	Select next word	Neither properties nor text input dialog open
Page Up	Apply word properties and select previous word	Properties or text input dialog open
	Select previous word	Neither properties nor text input dialog open
Ctrl + ⮌⮊⮉⮋	Select neighbour word	
Ctrl + F	Open search dialog	
Ctrl + T	Enable/disable text overlay	
1	Create initial word	
2	Split tool	
3	Split (cut) tool	
4	Contour detection (start from rectangle)	
5	Contour detection (start from polygon)	
S	Select components tool	
C	Create word from components	Components selected
R	Create rectangular word	
P	Create polygonal word	
I	Create isothetic polygonal word	
M	Merge selected words	
O	OCR selected words	Words selected
A	Adjust word outline to glyphs	
Ctrl + B	Converts the outlines of selected words to their bounding box	
Ctrl + I	Converts the outlines of selected words to isothetic format	
D	Auto detection tool	
B	Create text line from words (bottom-up)	

Glyph View

Shortcut	**Description**	**Scope**
F1	Switch to ‘Select’ tool	
F2	Switch to ‘Edit’ tool	
F10	Open dialog for word properties	

F11	Open text input dialog	
Ctrl + A	Select all glyphs	
Ctrl + C	Copy text content to Windows clipboard; Copy glyph(s)	Glyph(s) selected
Ctrl + V	Paste previously copied glyph(s) (activates tool with preview)	
Delete	Deletes the selected glyph(s)	Select or hand tool
	Deletes multiple polygon points	Edit tool + points selected
	Deletes a polygon point	Edit tool + mouse hover over polygon point
Page Down	Apply glyph properties and select next glyph	Properties or text input dialog open
	Select next glyph	Neither properties nor text input dialog open
Page Up	Apply glyph properties and select previous glyph	Properties or text input dialog open
	Select previous glyph	Neither properties nor text input dialog open
Ctrl + ⮜⮞⮝⮟	Select neighbour glyph	
Ctrl + F	Open search dialog	
Ctrl + T	Enable/disable text overlay	
1	Create initial glyph	
2	Split tool	
3	Split (cut) tool	
4	Contour detection (start from rectangle)	
5	Contour detection (start from polygon)	
S	Select components tool	
C	Create glyph from components	Components selected
R	Create rectangular glyph	
P	Create polygonal glyph	
I	Create isothetic polygonal glyph	
M	Merge selected glyphs	
O	OCR selected glyphs	Glyphs selected
Ctrl + B	Converts the outlines of selected glyphs to their bounding box	
Ctrl + I	Converts the outlines of selected glyphs to isothetic format	
B	Create word from glyphs (bottom-up)	

Dewarping View

Shortcut	Description	Scope
Delete	Removes selected grid	
1	Activates ‘Create grid’ tool	
2	Activates ‘Add vertical grid line’ tool	
3	Activates ‘Add horizontal grid line’ tool	
4	Activates ‘Initial grid’ tool for assisted grid creation	
F2	Switch to ‘Edit’ tool	

Administration

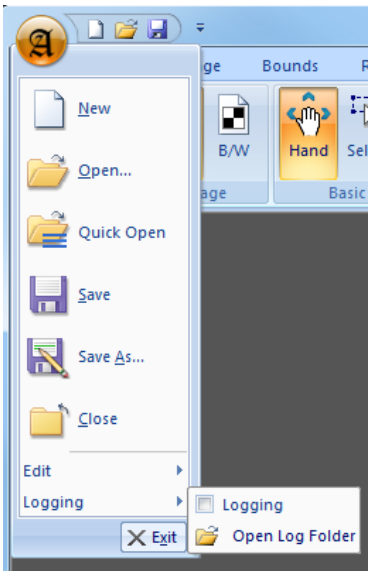
Location of User Defined Settings

The user can customise some settings in Aletheia. The settings are stored as an XML file. The location of this file can be changed by modifying the 'settingsFilePath' value within the aletheia.ini file.

- Open the file aletheia.ini (located in 'bin' within the main folder)
- Find the key 'settingsFilePath' of the 'main' section
- Change the value to one of the following settings:
    - APPDATA - Stores the settings in the application data folder of the current windows user (this is the default location)
    - ROOT - Stores the settings in the same folder as the Aletheia executable
    - TEMP - Stores the settings in the folder specified by the TEMP environment variable
    - *Path* - Stores the settings in the folder specified by the path (e.g. c:\temp)

Logging

For error tracing, all user actions can be logged to a text file. The logging can be enabled in the Aletheia menu:



Each time Aletheia is started, a new log file is created. The files are stored in the shared application data section of the current Windows user. To view the log files click 'Open log file folder' in the menu.

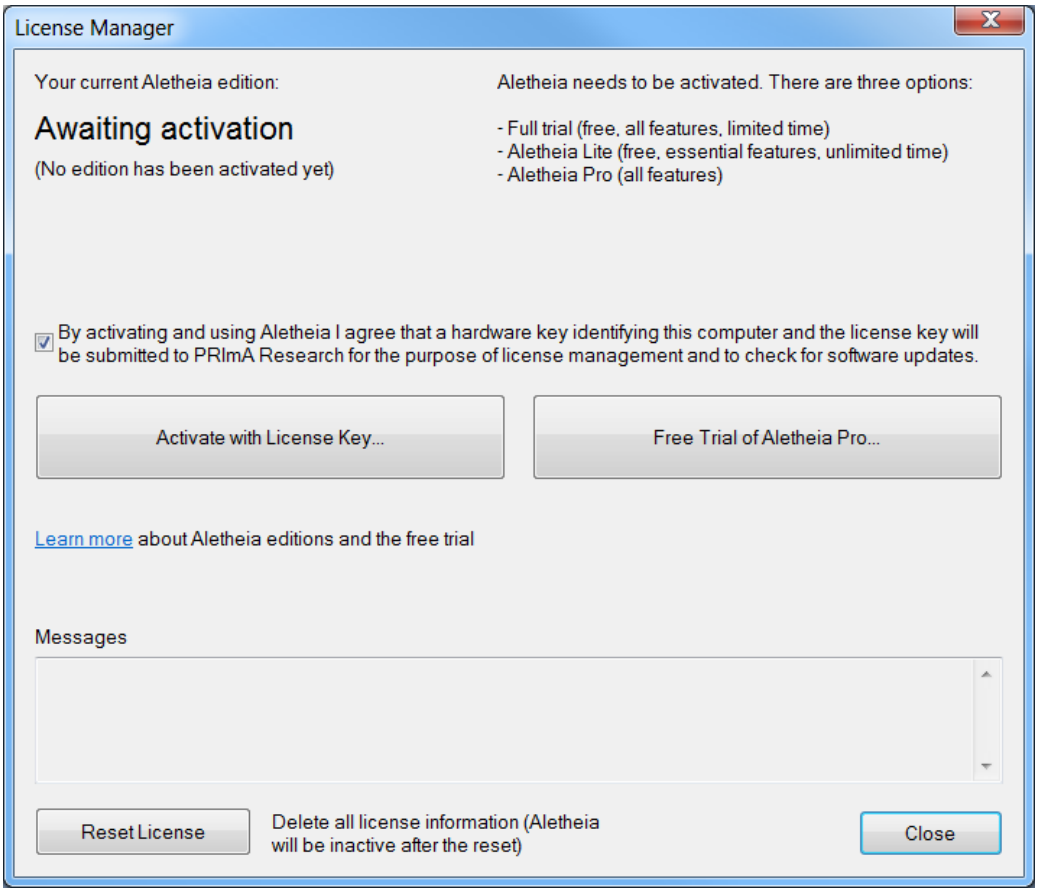
Note: Log files older than 24 hours will be deleted automatically.

Licence Management

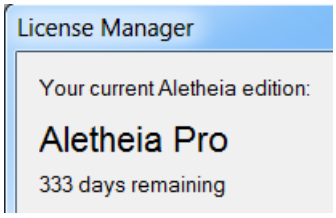
There are different Editions of Aletheia, all of which require activation. You can activate an edition using the "Licence Manager" dialog (Internet connection needed). If you run Aletheia for the first time the dialog should open automatically.

To open the Licence Manager:

- In the Aletheia menu click on "Licence Manager…"
  OR
- In the Welcome Window click on "Activate…" or "Manage…"



Once activated, you can see your current edition and the expiry date:

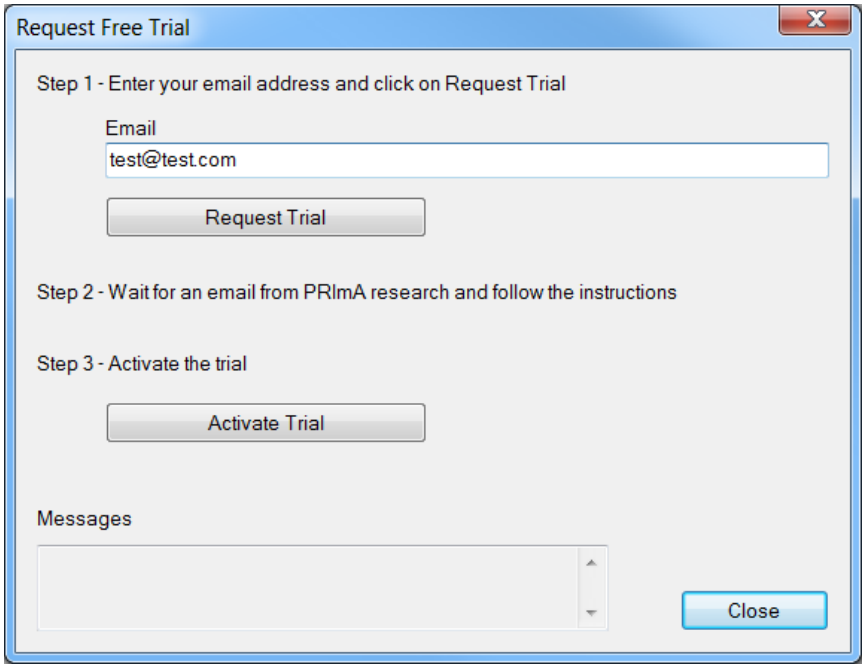


Activating using the Free Trial of Aletheia Pro

The free trial offers all features of the Pro edition for a limited amount of time.

To activate the trial (requires Internet access):

- Open the license manager
- (Agree to the license statement by ticking the box)
- Click on “Free Trial of Aletheia Pro…” (opens “Request Free Trial” dialog)
- Enter your email address
- Click on “Request trial”
- Wait for an email from PRImA Research and follow the instructions in the email to verify your email address
- Click on “Activate trial” once your email address is verified



Activating using a Licence Key

Visit the PRImA Research website to get a licence key for Aletheia Lite or Aletheia Pro:

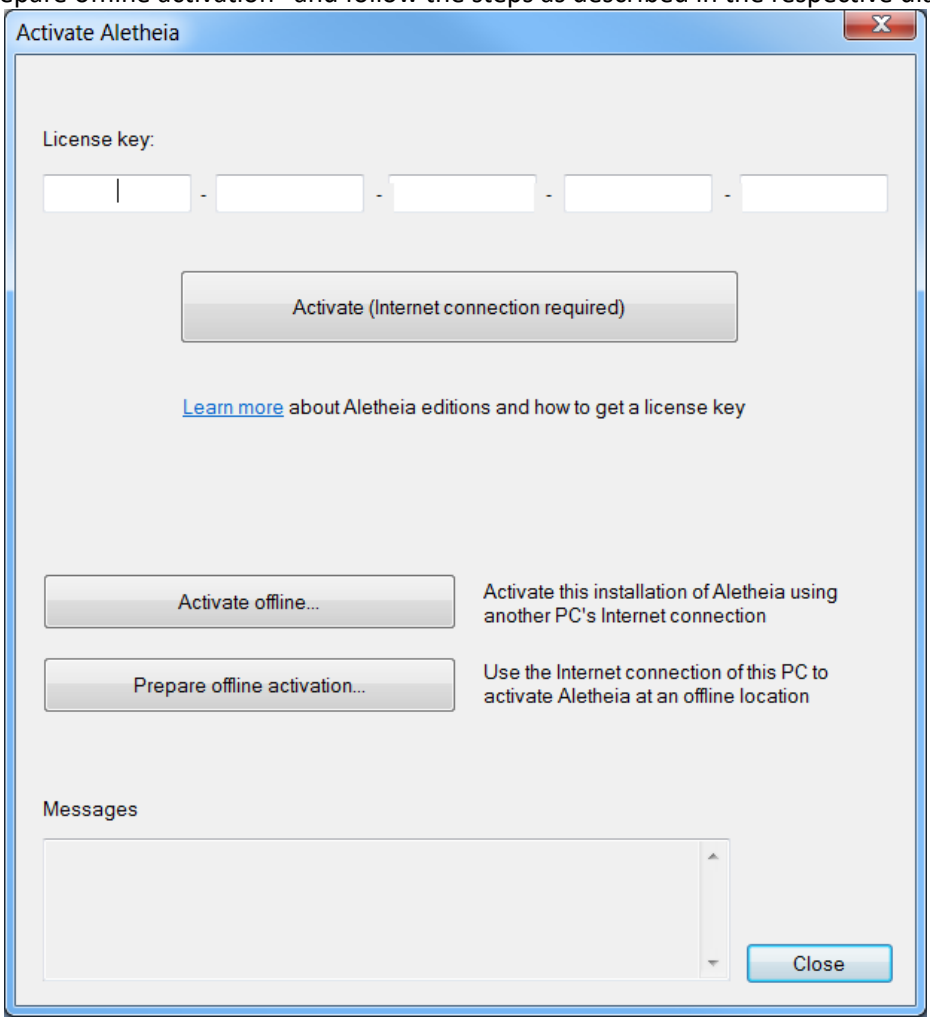
www.primaresearch.org/tools/Aletheia/Editions

To activate with a licence key (requires Internet access):

- Open the licence manager
- (Agree to the licence statement by ticking the box)
- Click on “Activate with licence key” (opens “Activate Aletheia” dialog)
- Enter your key
- Click on “Activate”

Note:

It is possible to activate Aletheia on a machine that has no Internet access. To do so use “Activate offline” and “Prepare offline activation” and follow the steps as described in the respective dialogs.



Licence Storage Location

Aletheia needs to be activated once via the Internet. All licence information is then stored locally.

It is possible to customise the location where Aletheia stores licence information:

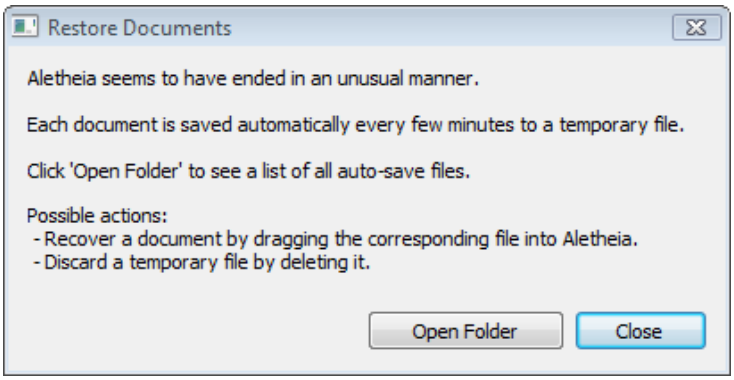
- Open the file aletheia.ini (located in ‘bin’ within the main folder)
- Find the key ‘licenseStorageLocation’ of the ‘main’ section
- Change the value to one of the following settings:
  - APPDATA
    - Stores the licence information in the application data folder of the current windows user (this is the default location)
    - Valid only for the current user
    - This is the default
  - ROOT
    - Stores the licence information in the same folder as the Aletheia executable
  - *Path*
    - Stores the licence information in the folder specified by the path (e.g. c:\temp)

Note: The current user has to have write permission for the specified folder, otherwise activating Aletheia will fail.

Error Messages and Warnings

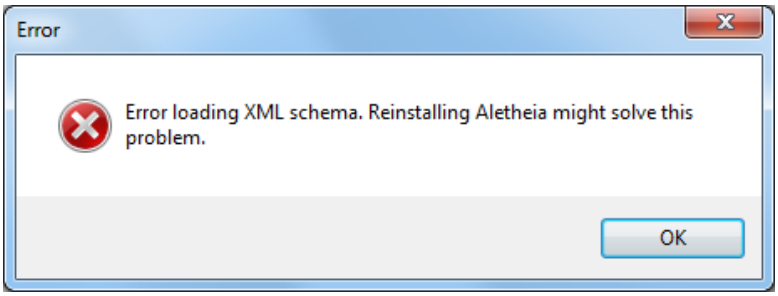
Messages on Starting Aletheia

Aletheia Seems to Have Ended in an Unusual Manner



This message is displayed, if there are auto-save files in the backup folder. Each document is regularly saved automatically to the backup folder. These temporary copies are deleted if Aletheia ends in a controlled way. If Aletheia ends unexpectedly through a program error, the auto-save files can be used to recover the last version of document files.
To open auto-save files, click ‘Open Folder’ and drag & drop the files to recover into Aletheia. Once open you can check the content of the files and save them under their original name.

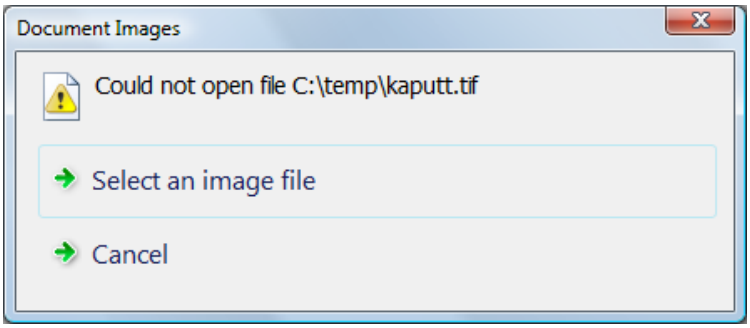
Error loading XML schema



This message indicates that the XML schema for PAGE files could not be loaded. Schema files are located within the Aletheia installation folder: <Aletheia>\schema\...
Try to reinstall Aletheia to solve this problem.

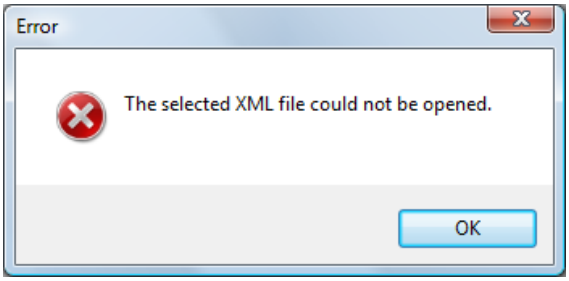
Messages on Creating or Opening a Document

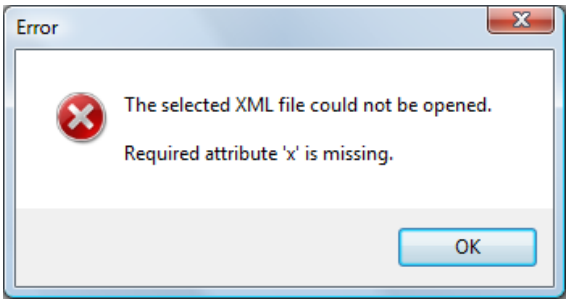
Could not open File



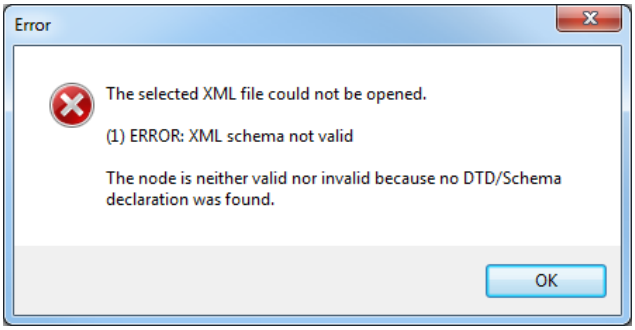
This message is displayed, if the document image file is corrupt. Check the image with an external tool for correctness or select another image.

XML File could not be opened



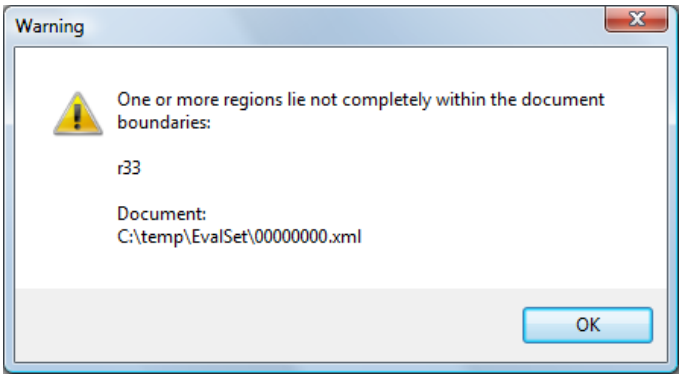


Means that the selected XML file is not in PAGE format. Make sure you select a PAGE document layout file.
If the error persists, check the file in an external editor.
If the document doesn't comply with the XML schema, error details are shown within the message.



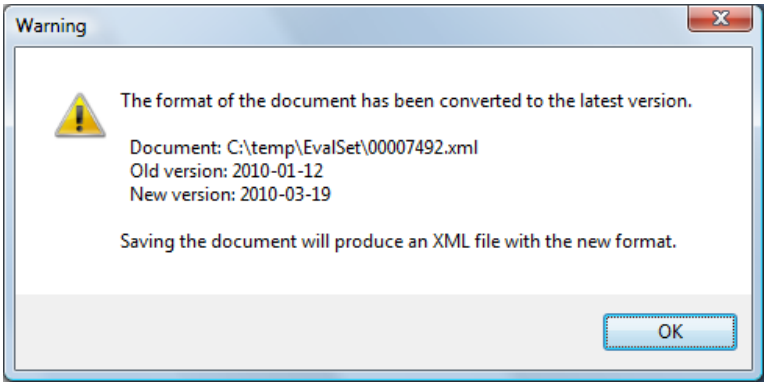
This message indicates that the XML schema file could not be found. This can happen when the schema file is located on the web and the computer is not connected to the internet.

Regions out of document boundaries



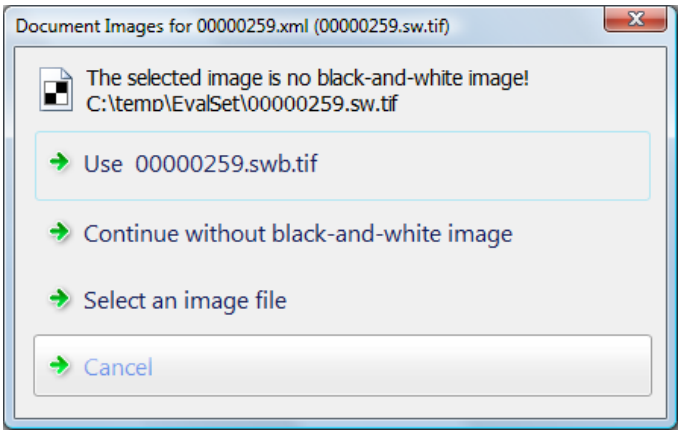
This message indicated that some layout regions have outline points that lie beyond the image dimensions (less than zero or greater than width/height). The involved regions are listed within the message. You can also use the validator (Tools->Validate Document) to locate problematic regions.

Document Format Conversion



This message indicated that the XML format of the opened document was out of date and has been converted to the recent schema. This message will not be displayed again, once the document has been saved again.

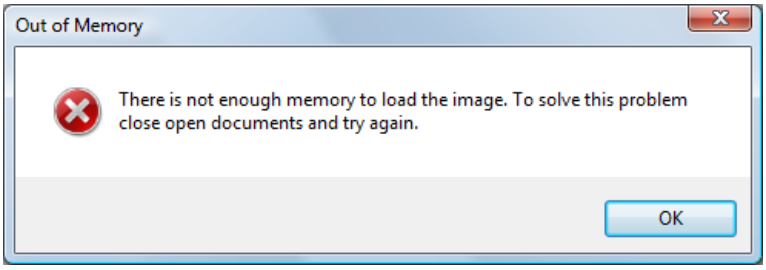
The Selected Image is no Black-and-White/Colour Image



This message is shown when an image file is selected, that is not of the required type. Aletheia detects the image type using the 'Bits per Pixel' header entry. If the value is one bit per pixel, the image is assumed to be black-and-white. Otherwise it is assumed to be a colour or grey level image.

Either click ‘Select an image file’ and choose another file (make sure it has the correct colour depth) or click ‘Continue without black-and-white image’/’Continue without colour image’ to open the document anyway. If you choose the second option, some features won’t be available in Aletheia.
If the problem persists, the image header may be damaged. In that case check the image with an external tool.

Out of Memory



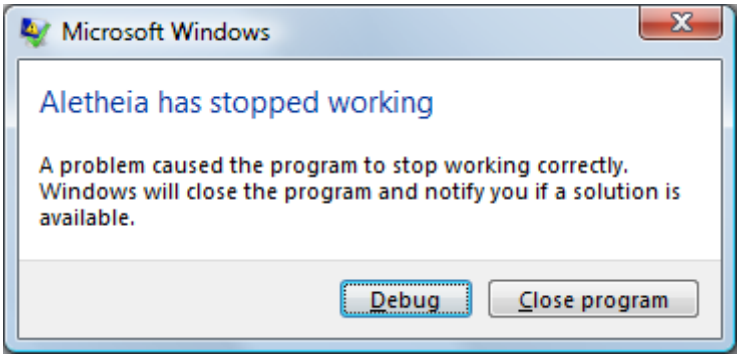
This message means that there is not enough memory to open the document. Close documents that needn’t be open necessarily and try again.
However, we recommend to restart Aletheia after this message, because after-effects may occur.

Aletheia Expires / has Expired

If you have a limited version of Aletheia, there is an expiry date. Open the ‘About’ dialog in the ‘Help’ menu, to check the expiry.
Aletheia shows a message on start-up, if less than 5 days of usage are left. If Aletheia has expired, Aletheia stops after confirming the message box.

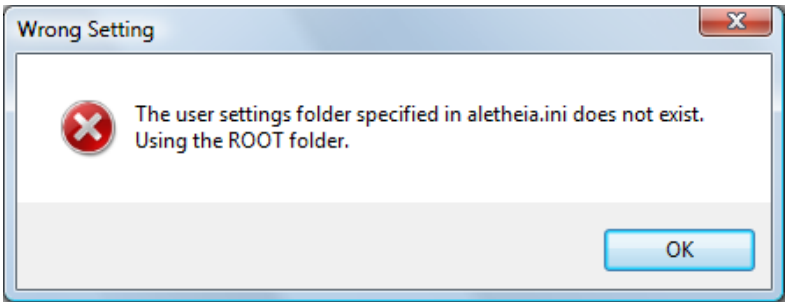
General Error Messages

Aletheia has Stopped Working



Internal program error. Lost data may be recovered using the ausosave files.

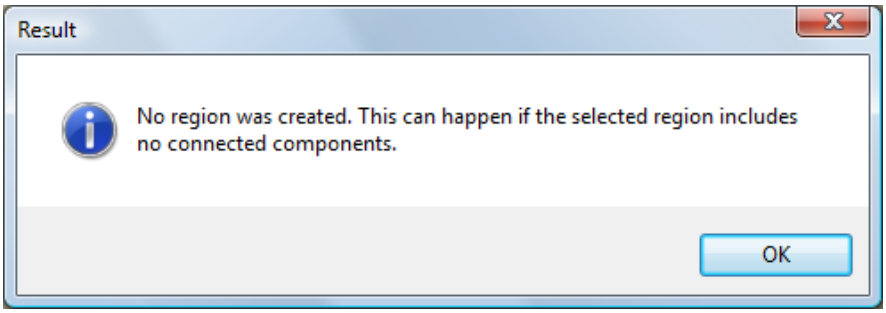
Wrong Setting



This message indicates that the specified folder for the user settings doesn't exist. This message may also occur, if the Aletheia.ini file is missing. Check if the file exists and check if the specified user settings folder exists.

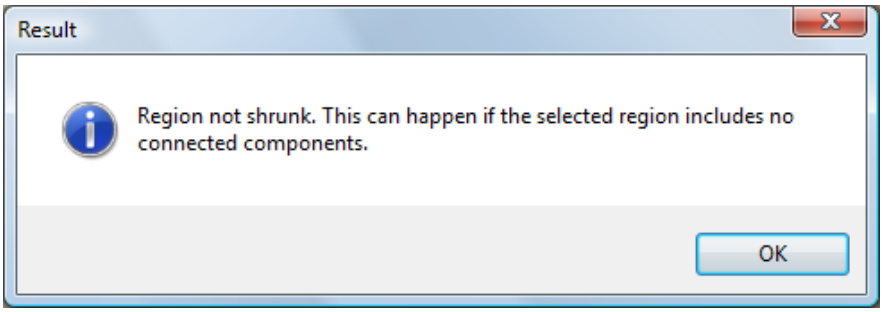
Tool Messages

No Region Created



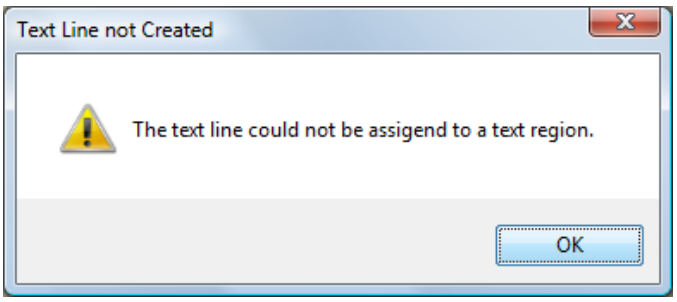
This message indicates that the text region shrinking algorithm didn't succeed. This is usually the case, when the drawn region doesn't contain connected components. Switch to the black-and-white image, to have a better view of the connected components.

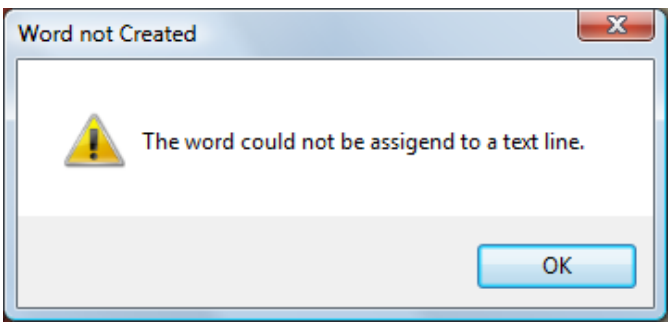
Region not Shrunk

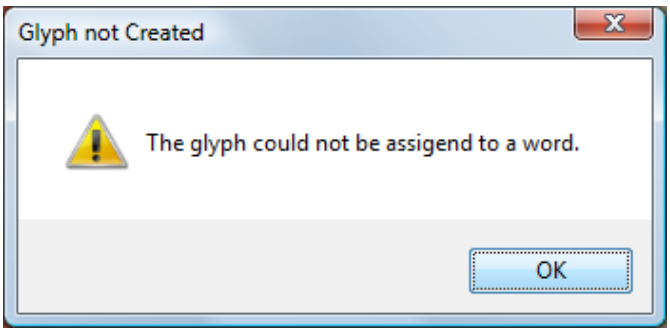


This message indicated that the component shrinking algorithm didn't succeed. This is usually the case, when the drawn region doesn't contain connected components. Switch to the black-and-white image, to have a better view of the connected components.

Text Line not Assigned to Region, Word not Assigned to Text Line, ...







These messages indicate, that a new text line (word, glyph) could not be assigned to a parent region (text region, text line, word), because no such region was found at the position of the new region.

Credits

PRImA Research Lab
University of Salford
United Kingdom

www.primaresearch.org

Director of research group	*Apostolos Antonacopoulos*
Aletheia project lead	*Christian Clausner*
Design	*Christian Clausner* *Christos Papadopoulos* *Stefan Pletschacher*
Development	*Christian Clausner*
Testing	*Christian Clausner* *Christos Papadopoulos* *Stefan Pletschacher*
Documentation	*Christian Clausner*

Copyright Notes for Third Party Extensions

LibTIFF

OpenCV

Tesseract

Leptonica is required. (www.leptonica.com).

Leptonica (used by Tesseract)

This work is licensed under a Creative Commons Attribution 2.5 License.

Copyright (C) 2001 Leptonica. All rights reserved.

Redistribution and use in source and binary forms, with or without
modification, are permitted provided that the following conditions
are met:
1. Redistributions of source code must retain the above copyright
   notice, this list of conditions and the following disclaimer.
2. Redistributions in binary form must reproduce the above
   copyright notice, this list of conditions and the following
   disclaimer in the documentation and/or other materials
   provided with the distribution.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS
``AS IS'' AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT
LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR
A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL ANY
CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL,
EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO,
PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR
PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY
OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING
NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS
SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

Dejavu Font

Fonts are (c) Bitstream (see below). DejaVu changes are in public domain.
Glyphs imported from Arev fonts are (c) Tavmjong Bah (see below)

Bitstream Vera Fonts Copyright

Copyright (c) 2003 by Bitstream, Inc. All Rights Reserved. Bitstream Vera is a trademark of Bitstream, Inc.

Permission is hereby granted, free of charge, to any person obtaining a copy of the fonts accompanying this license ("Fonts") and associated documentation files (the "Font Software"), to reproduce and distribute the Font Software, including without limitation the rights to use, copy, merge, publish, distribute, and/or sell copies of the Font Software, and to permit persons to whom the Font Software is furnished to do so, subject to the following conditions:

The above copyright and trademark notices and this permission notice shall be included in all copies of one or more of the Font Software typefaces.

The Font Software may be modified, altered, or added to, and in particular the designs of glyphs or characters in the Fonts may be modified and additional glyphs or characters may be added to the Fonts, only if the fonts are renamed to names not containing either the words "Bitstream" or the word "Vera".

This License becomes null and void to the extent applicable to Fonts or Font Software that has been modified and is distributed under the "Bitstream Vera" names.

The Font Software may be sold as part of a larger software package but no copy of one or more of the Font Software typefaces may be sold by itself.

THE FONT SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT OF COPYRIGHT, PATENT, TRADEMARK, OR OTHER RIGHT. IN NO EVENT SHALL BITSTREAM OR THE GNOME FOUNDATION BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, INCLUDING ANY GENERAL, SPECIAL, INDIRECT, INCIDENTAL, OR CONSEQUENTIAL DAMAGES, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF THE USE OR INABILITY TO USE THE FONT SOFTWARE OR FROM OTHER DEALINGS IN THE FONT SOFTWARE.

Except as contained in this notice, the names of Gnome, the Gnome Foundation, and Bitstream Inc., shall not be used in advertising or otherwise to promote the sale, use or other dealings in this Font Software without prior written authorization from the Gnome Foundation or Bitstream Inc., respectively. For further information, contact: fonts at gnome dot org.

Arev Fonts Copyright

ALTO XML Schema

See http://www.loc.gov/standards/alto/

ALTO: Analyzed Layout and Text Object

Originally created during the EU-funded Project METAe, the Metadata Engine Project (2001 - 2003), by Alexander Egger (1), Birgit Stehno (2) and Gregor Retti (2), (1) University of Graz and (2) University of Innsbruck, Austria with contributions of Ralph Tiede, CCS GmbH, Germany. Prepared for the Library of Congress by Ralph Tiede, CCS GmbH, with the assistance of Justin Littman (Library of Congress).

MUPDF

The mupdf tool is used to convert PDFs to images.
See https://mupdf.com/

From the README:
ABOUT

MuPDF is a lightweight open source software framework for viewing and converting
PDF, XPS, and E-book documents.

See the documentation in docs/index.html for an overview.

Build instructions can be found in docs/building.html.

LICENSE

REPORTING BUGS AND PROBLEMS

The MuPDF developers hang out on IRC in the #mupdf channel on irc.freenode.net.

Report bugs on the ghostscript bugzilla, with MuPDF as the selected component.

http://bugs.ghostscript.com/

If you are reporting a problem with a specific file, please include the file as
an attachment.