How To Process Tables in Transkribus

This is a short introduction to marking up and exporting tables as well as to the semi-automatic table processing using Transkribus and nomacs. Segmenting printed or hand-drawn tables using the Table Editor in Transkribus will add graphical lines into your image and assign a tabular structure to the layout of your documents. It also allows you to export your transcriptions as a Microsoft Excel spreadsheet. This guide applies to images in one Transkribus document following the same table print or template.

Download the Transkribus Expert Client, or make sure you are using the latest version:
- https://transkribus.eu/

Consult the Transkribus Wiki for further information and other How to Guides:
- https://transkribus.eu/wiki/

Transkribus and the technology behind it are made available via the following projects and sites:
- https://read.transkribus.eu/
- https://transcriptorium.eu/
- https://github.com/transkribus/

Contact
- The Transkribus Team: email@transkribus.eu
Contents

Introduction ........................................................................................................................................... 3
Segmentation ....................................................................................................................................... 3
   Create text regions .......................................................................................................................... 3
   Create the table .............................................................................................................................. 3
   Segment the table ........................................................................................................................... 3
   Correct the cells in the table .......................................................................................................... 3
Add graphical information .................................................................................................................. 5
   Border Mark-up ............................................................................................................................. 6
Add baselines ...................................................................................................................................... 6
Correct baselines stretching more than one cell ................................................................................. 7
Transcribe the table headers .............................................................................................................. 8
Example results of table markup ...................................................................................................... 8
Transcription ....................................................................................................................................... 8
Export Tables ...................................................................................................................................... 9
Semi-automatic table processing ....................................................................................................... 9
   Create the template ........................................................................................................................ 9
Download the Document .................................................................................................................. 10
Delete non-template XML files ........................................................................................................ 10
Run the template matching process ................................................................................................ 10
Upload your documents into Transkribus .......................................................................................... 13
Add cell segmentation ...................................................................................................................... 14
Further processing ............................................................................................................................. 14
Credits ................................................................................................................................................ 14

The READ project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 674943.
Introduction

Printed and hand-drawn tables are common in historical documents of all types. Such tables can be marked up in Transkribus, either as the first stage in creating training data for Automated Text Recognition or simply to ready the documents for manual transcription.

Currently, tables must be manually drawn using the Table Editor in Transkribus. Technology which will allow the automated recognition of tables is under development and will be made available to users soon.

Often multiple pages follow the same table print or table template, so the table mark-up only has to be done for the first occurrence of the same print and can be distributed to the remaining pages using the nomacs toolkit.

The first section of this guide describes the manual creation of a table structure in Transkribus and the transcription of the text it contains. The second section gives instructions on working with table templates, which were created in Transkribus, and how to apply them to several pages using a method called batch processing in the nomacs tool.

Finally, this document also explains how tables can be exported for further data processing in standard spreadsheet tools.

Segmentation

Create text regions

- First, create text regions for any information not belonging to the table. This refers to information at the top, bottom or the sides of the page which is clearly not part of the table such as:
  - Page numbers
  - Line numbers
  - Dates
  - Any other markings or annotations
- For more information on creating text regions, see the “Segmentation” section of How To Transcribe Documents with Transkribus – Introduction.

Create the table

- Select the “Add other item” button in the Canvas menu and then click “Add a table”
- Click on the top left corner of the table in the image and then click on the bottom right corner

Segment the table

You can now segment your table into rows and columns
How to Process Tables in Transkribus

- To begin, make sure you are in “Selection mode”. Press the “ESC” key on your keyboard or click the “Selection mode” button in the Main menu.
- Click on the table region that you have created.
- To create rows, click the “Splits a shape with a horizontal line” button in the Canvas menu.
- Move your cursor across the page and click wherever you want to create a horizontal line.
- To create columns, click the “Splits a shape with a vertical line” button in the Canvas menu.
- Move your cursor across the page and click wherever you want to create a vertical line.
- Continue until all table cells are marked.

**Note:** Depending on the layout of your table, you might want to treat the spine of the book like an extra column (as in Figure 1). You can also mark-up this column on table cell level using the “book-binding” tag in the “Metadata/structural” tab.
How to Process Tables in Transkribus

Correct the cells in the table
In some cases, it may be necessary to merge cells together in order to reflect cells spanning multiple rows or columns.

- Make sure you are in “Selection mode” by pressing the “ESC” key on your keyboard or clicking the “Selection mode” button in the Main menu
- To select cells to merge, hold down the “CTRL/CMD” key on your keyboard and then click on the relevant cells in your table
- Click the “Merges the selected shapes” button in the Canvas menu
- Continue with all cells until the expected structure is achieved. In the example below, merging must be completed for each of the highlighted set of cells.

If you focus on having the perfect table segmentation, it may also be necessary to correct the shapes of some of the cells in your table. The segmented green lines should then correspond to the lines of your table as far as possible. In order to do so,

- Select the table cell you wish to edit
- Click and drag the big green dots to move the position of the lines
**Note:** For export and automatic processing, having straight, rectangular lines close to the original table borders is perfectly sufficient.

### Add graphical information

**Border Mark-up**

Cell borders (graphical lines) need to be marked when they are visible.

- Right-click on the cell that you wish to mark up
- Click “Mark-up borders” in the pop-up menu or use the button to open the border markup menu
- Choose the correct options to describe the border of the cell

**Note:** You can choose multiple cells at once by choosing “Select all cells” or “Select row cells”. Selecting or de-selecting cells works by holding down the command (Ctrl) key and clicking on a different cell.

### Add baselines

The next step is to add baselines to your table. The baselines should reflect the logical flow of text and can therefore run over the cell borders if necessary.

- You can either draw the baselines by hand or use the automatic baseline detection tools in Transkribus

**Note:** the line finding tool created by the [Computational Intelligence Technology Lab](https://www.citlab.uni-rostock.de/) at Universität Rostock is currently the most effective for the automatic recognition of baselines in tables. In the “Layout Analysis” section of the “Tools” tab click “Method: CITLab Advanced”.
If you detect baselines automatically, you may need to correct the generated lines or move them to the correct cell. You may also want to check the reading order and correct your baselines. For more information on adding and correcting baselines, see the “Segmentation” section of How To Transcribe Documents with Transkribus – Introduction.

Correct baselines stretching more than one cell

You may find that the automatic layout tool on table cells strictly obeys the cell borders. Baselines stretching multiple cells are divided. You can use the merging tool to combine those partial baselines. In case you want to merge baselines stretching more than one cell, move them first to the same cell, select them and use the merge tool.

- Open the “Layout tab”.
- Click on the first cell in the image where your baselines should be placed. This will highlight the respective position in the structure tree.
- Expand the arrows to display the Line elements.
- Select the lines you want to move from the tree holding down the Ctrl key.
- Drag the Lines to the correct cell.
How to Process Tables in Transkribus

- Use the merge tool to fix the layout of the lines.

Transcribe the table headers
Especially for given forms or tables, the headers remain the same over several pages. Any transcribed information contained in the table template will automatically be carried over by the table-matching tool.

Example results of table markup

Transcription
- Transcribe the text of your table exactly as it appears in the image.
- Click on a cell in your table to start transcribing and then move through the other cells in your table.
- If you are transcribing text as training data for Automated Text Recognition, the reading order of your transcription is not important.
- If you are transcribing text for research purposes, you may want to adjust the reading order of the baselines.
- You can also run an automated text recognition (HTR model) on your segmented document. For more information see How To Train A Handwritten Text Recognition Model in Transkribus.
In the next section of the guide, you will find out how to create a table template that can recur across several images in your document.

Export Tables

Once you have segmented and transcribed a page, you can export the results of your transcribed tables into XLS format.

- Click the “Export Document” button in the Main menu
- At the top of the box, select the location where you would like your exported files to be saved
- In the “Choose export format” section, select “Table Export into Excel”
- At the bottom right of the box, make sure that you select the number of pages that you wish to export.
- Export a single page

**Note:** Only tables and their contents are exported, text regions will be ignored. If your selection of pages contains no tables, Transkribus will show you an error message and stop the export process.

Semi-automatic table processing

Usual book layouts show many consecutive pages using the same layout. You can use the nomacs toolkit to transfer one template page of segmentation (table borders and header baselines including header transcriptions) to subsequent pages.

Create the template

In order to use the semi-automated methods, follow the segmentation section of this document to mark up the basic structure of your table on the first page.

Your template page (usually the first page showing the typical table layout and structure) should contain the following elements:

- Segmented table structure (table area, table header cells, table rows only if they are not manually drawn)
- Graphical information, i.e. visibility of table cell borders
- Baselines for any lines which remain unchanged on every page of your document
- Transcription for any text which remains the same on every page such as headers. Any transcribed information contained in the table template will automatically be carried over by the table-matching tool.

**Note:** With the existing version of the table matching tool, only one table per page is allowed. This restriction will be discarded in the future.

**Note:** The template matching is based on the recurrent layout of the table. Thus, e.g. manually drawn row separators can result in a different number of rows and should not be marked. You can run automatic layout analysis only on the header cells. You can do so by right-clicking on the first
header and selecting “Select row cells” from the context menu. Untick the “Find Text Regions” in the “Layout Analysis” section of the “Tools” tab and the baselines of the selected cells will be found.

**Note:** If your document contains hand-drawn table cell borders, we recommend to leave the non-header table rows in the template as one large, artificial table row. The number of rows on each page might vary, so this influences the table matching process and quality. Instead, you should add horizontal lines to your tables after running through the batch processing.

**Download the Document**

In order to work with the nomacs tool, you need to export your Transkribus document first.

- Open the export tool
- Choose “Transkribus Document” as export format
- Tick “Export Page” and “Export Image”
- Choose the “filename” option in the “Filename pattern” field

**Note:** In order to synchronize the results of your batch processing back to Transkribus, the filenames must remain unchanged.

**Delete non-template XML files**

- Browse to the location to which you just downloaded your Transkribus document
- Open the “page” folder
- Delete all but your template file – hint: check the file sizes and delete all but the largest file

**Run the template matching process**

- Open nomacs ReadFramework

  - Click on the  button to open the folder where your Transkribus export is located

1 Download and install the latest release version from [https://github.com/TUWien/ReadModules/releases](https://github.com/TUWien/ReadModules/releases)
- Display the contents of the graphical overlay by choosing “Plugins”  \textgreater \textgreater “PAGE Visualization”

- Go to “Tools” \textgreater “Batch processing”
- Click on “Input”, enter the path to your input folder in the text box and choose every image except your template image by holding the “CTRL/CMD/Shift” key while clicking on the
How to Process Tables in Transkribus

Afterwards you have to press the “Select File” Button. You can also see the list of selected files in the “Files List” tab. Here you can also edit the file list.

- Click on “Plugins” and select “Forms Analysis” → ”Apply template (Match)” and adjust the parameters.
- Make sure to enter the correct path for the form template.
  
  **Hint**: you can copy and paste the file path and file name from the right side “Page
How to Process Tables in Transkribus

Visualization” panel just above “Drag XMLs here” field. Make sure to adjust the filename to perfectly match your template.

- Adjust the template matching parameters
  Good settings are currently “variationThresholdLower” at 0,4 and “variationThresholdUpper” at 0,6

**Note:** The threshold parameters allow the table matching to be more “flexible”; this means that rows or columns can have varying width, etc. A value of 0 means that the row/column width and height must match perfectly with the template. Higher values allow a greater variation.

- Select the “Output Directory” where you want nomacs to store your results
- Hit the play symbol to start the batch processing
- Move all newly created .xml files to the “page” folder. This is needed for the upload into Transkribus.

**Note:** It might be necessary to adjust some parameters and re-run the template matching process in the nomacs tool, so make sure to have a look at the results by browsing through the files in nomacs.

Upload your documents into Transkribus

At the time of writing, nomacs and Transkribus are not directly connected, so you have to synchronize your new XML files back to Transkribus.
How to Process Tables in Transkribus

You can easily do so by clicking ☰ in the Main menu and choosing “Document”→“Sync local transcriptions with doc...” in Transkribus.

Add cell segmentation
Once you have finished synchronizing your table layout, you can add the remaining layout structures and correct table borders where necessary.

- Cut the large table body into the cells as they are drawn on the original image using the horizontal cutting tool.
- Select all table cells from the context-menu on the table.
- Hold down the “CTRL/CMD” key and untick every cell which already contains correct and transcribed baselines, i.e. unselect the header cells which are part of your template.
- Run the automatic layout analysis only on the selected cells by unticking the “Find Text Regions” option in the “Tools” tab.

Further processing
Once your table segmentation and table cell segmentation is finished, you can add baselines and transcriptions. You can also export your table into Excel format. For further details, check the corresponding sections of this guide.

Credits
We would like to thank the many users who have contributed their feedback to help improve the Transkribus software.

Transkribus is made available to the public as part of the H2020 e-Infrastructure Project READ (Recognition and Enrichment of Archival Documents) which received funding from the European Commission under grant agreement No. 674943.