READ
Recognition and Enrichment of Archival Documents

How To Transcribe Documents with Transkribus-Simple Mode

Version v1.3.8.1-Snapshot /(07_12_2017)

Transkribus is a platform for the automated recognition, transcription and searching of historical documents, using Handwritten Text Recognition (HTR) technology.

Transcripts generated with Transkribus can be used to:

- train a neural network ("model") which is capable of automatically recognising printed or handwritten documents
- create scholarly reliable and highly standardized transcripts which may serve as the basis for digital editions of documents.

This introduction enables you to quickly create training data for the automated recognition of your specific documents.

If you have already transcribed documents available – please consult our HowToUseExisitingTranscriptions paper.

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

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

Introduction

This guide explains the process of using Transkribus to create transcripts that can be used as training data for a Handwritten Text Recognition HTR model.

There is a simple three-step process for transcribing a document in Transkribus:

Step 1: Uploading

- Upload your documents to the Transkribus platform

Step 2: Segmentation

- Run the automated segmentation tool to create baselines for your document.

Step 3: Transcription

- Transcribe the text in the segmented lines.

Upload documents to Transkribus

- In order to be able to run the necessary tools on your documents they need to reside on the Transkribus server. This means that you **need to upload them to Transkribus.**
    - Note: **All collections and documents in Transkribus are private**. Only users authorised by you are able to see your documents. They are not made available to the public. Uploading a document to the Transkribus server is therefore a purely technical process.
- To upload click on the “Import Documents” button in the Main menu.

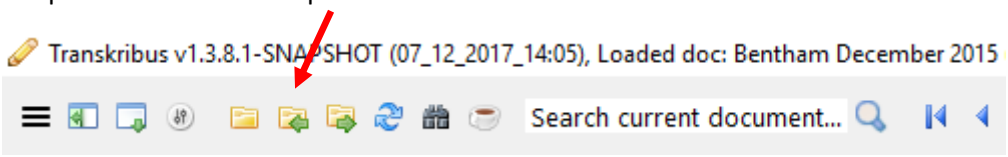


Figure 1 Upload files to your personal collection

- You have three options:
    - **Upload single document** from a local folder:
        - This option allows you to upload documents up to 500 MB
    - **Upload via FTP**
        - This is suitable if you want to upload several large documents
    - **Upload via URL** of DFG Viewer METS
        - This allows you to upload documents directly from repositories which support the DFG (Deutsche Forschungsgemeinschaft – German Science Funds) Viewer
    - **Extract and upload images from PDF**

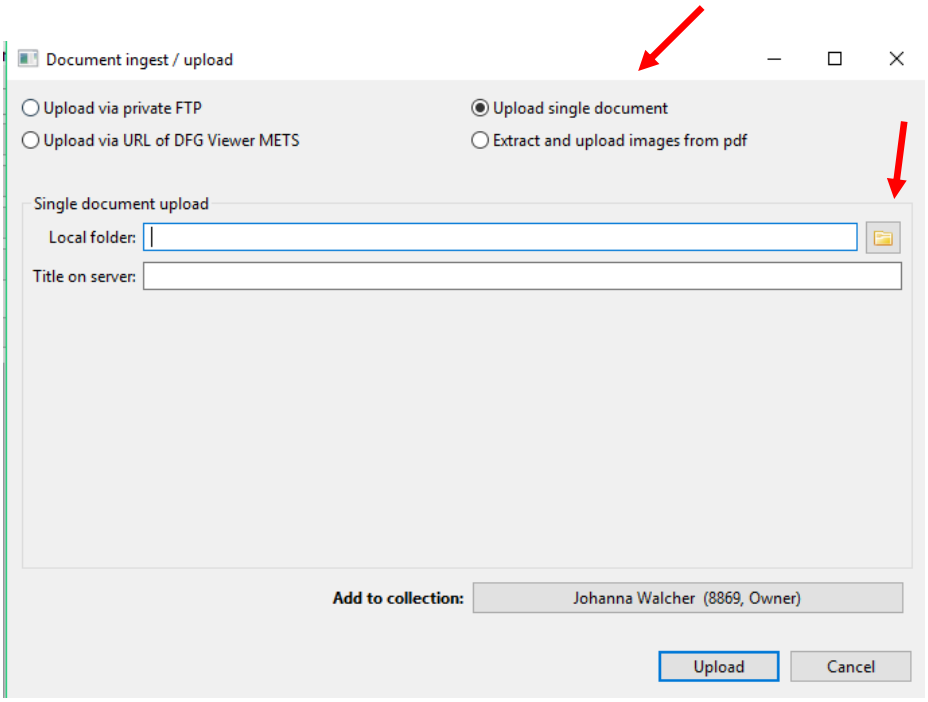


Figure 2 Select "Upload single document" for documents up to 500 MB

Segmentation

Introduction

- Once you have uploaded your documents to Transkribus, you are ready to start segmentation.
- For the “Handwritten Text Recognition” to work, the **text and image need to be connected** in Transkribus. This is achieved by dividing your documents into text regions, lines and baselines.

Viewing profiles

- Viewing profiles are available to help you with the tasks of segmentation and transcription.
- You can select between viewing profiles for “**Segmentation**” and “**Transcription**” by clicking the “Profiles” button in the Main menu.
- The “Segmentation” profile means that baselines are displayed in red, making it easier to spot any errors resulting from the automated segmentation process.
- The “Transcription” profile means that the text editor field will be displayed, allowing you to transcribe your document.
- Of course you can simply use the “default” profile to perform either task.

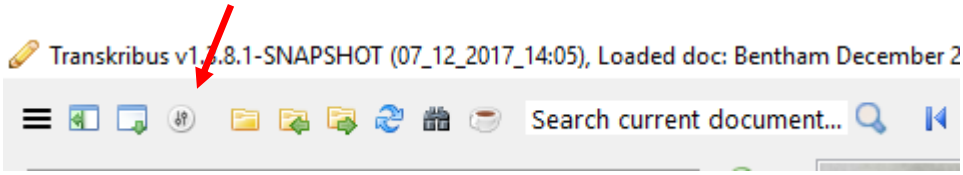


Figure 3 Viewing profiles for segmentation and transcription tasks

Automatically detect text regions, lines and baselines

- Select the “Segmentation” viewing profile from the Main menu.
- Select the “Tools” tab on the left side of the screen and go to the “Layout Analysis” section.
- Under “Method:” select “CITlab Advanced”.
- Select if you would like to run the layout analysis only for the current page, for distinct pages, or for the whole document.
- Make sure you select both, “Find Text Regions” and “Find Lines In Text Regions”.
- Click the “Run” button.

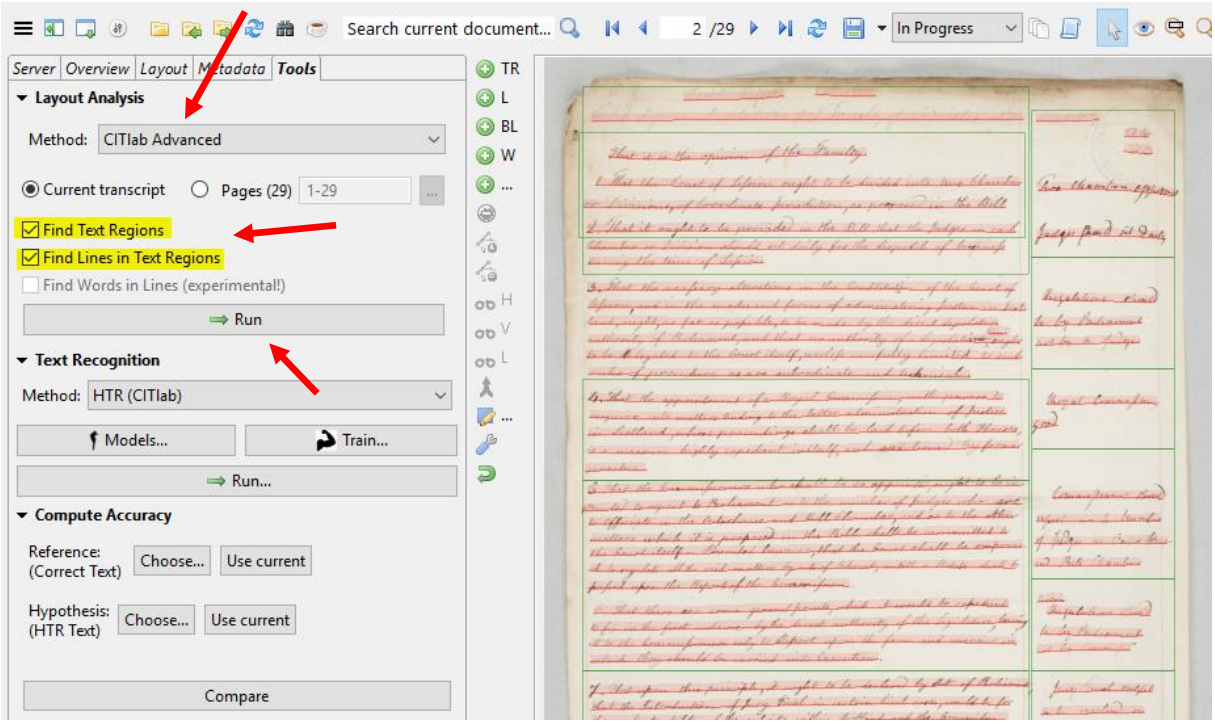


Figure 4 Perform automated segmentation in the “Tools” tab

- The **program will automatically detect** the text regions, lines and baselines in your document.
- In most cases, this automated process will work to a high level of accuracy.
- Where documents have a very complicated layout, you may need to perform some manual corrections.

Correcting the results of automated segmentation

- Note: if you are training a “Handwritten Text Recognition” model, the position of text regions does not need to be completely exact and the reading order of the text is not relevant.
- If you are working on a scholarly edition where a higher degree of accuracy is required, it is possible to manually correct the text as in the examples below.

A line has been missed or added by mistake

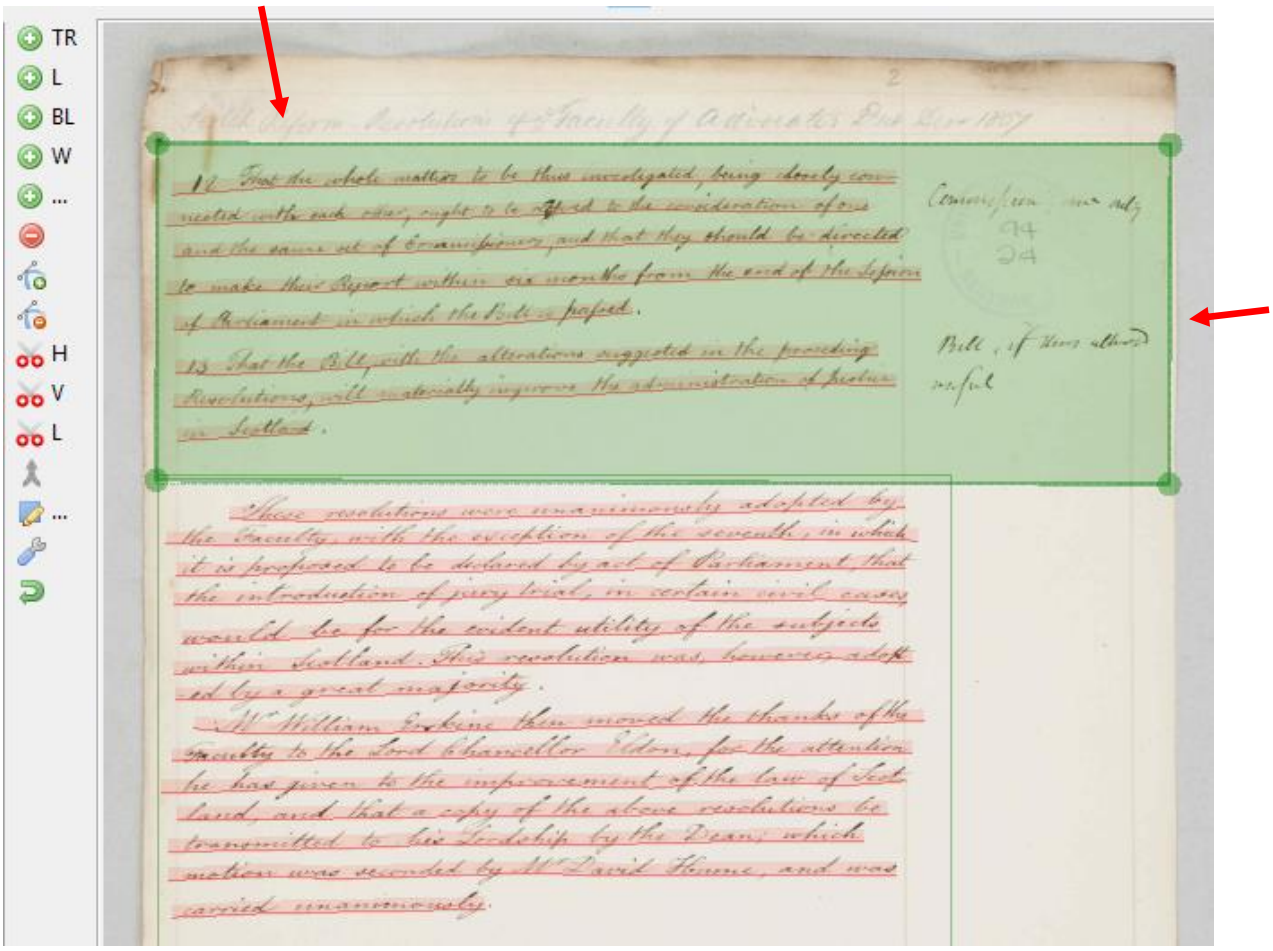


Figure 5 Add a line to an existing text region

- In the example above the first line had been missed by the program. If you would like to add it to the existing text region:
  - Click inside the region so that it is highlighted.
  - Drag the border of the text region as needed.

A marginal note needs to be split into a separate text region

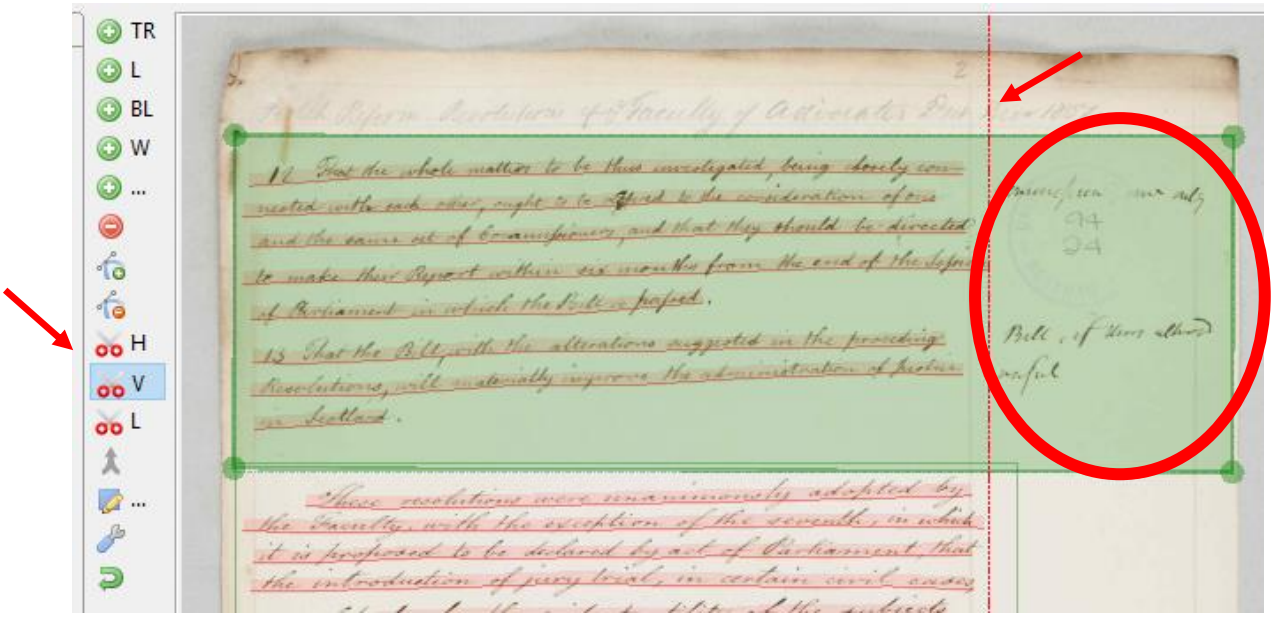


Figure 6 Split a text region

- If you need to split one region into two, you can do this with buttons in the Canvas menu.
- As shown in Figure 6, the "H-button" splits a text region horizontally.
- The "L-button" allows you to split a text region with customisable line.

Remove a region which is not needed

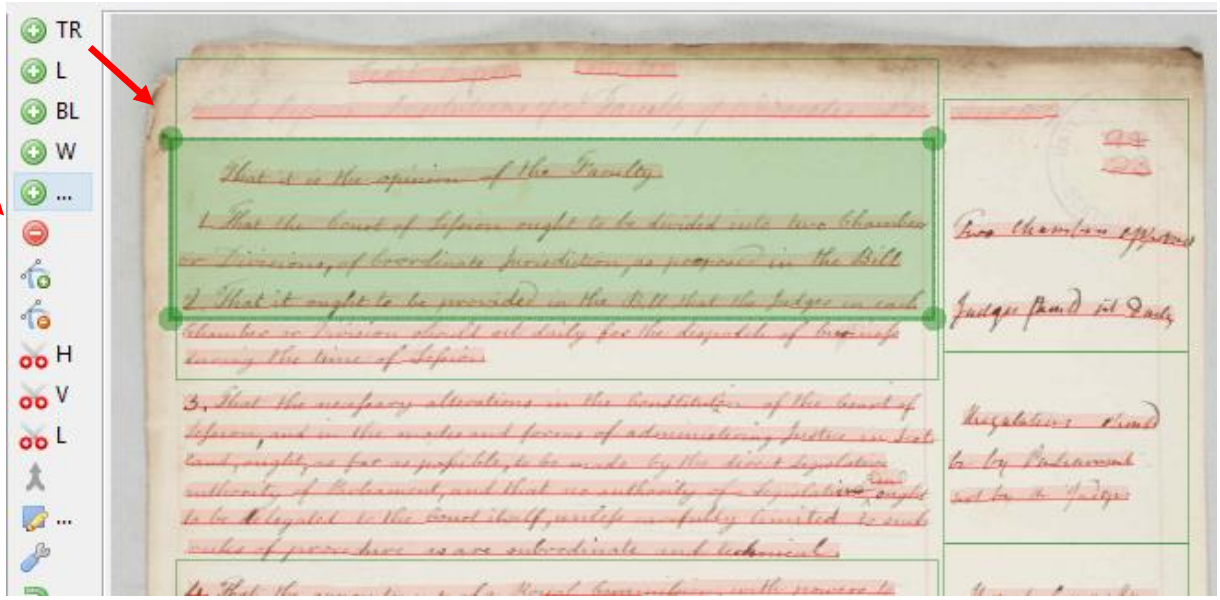


Figure 7 Remove region

- In the example above two regions are overlapping, so one can be deleted.
- Click on the text region you wish to delete, and click the red "Remove a shape" button.

Merge two regions

- Sometimes the program creates two text regions where only one is needed. In this case you can easily merge the two together.
  - Hold down the "CTRL" button on your keyboard and click on both text regions.
  - Click the "Merges the selected shapes" button in the Canvas menu.

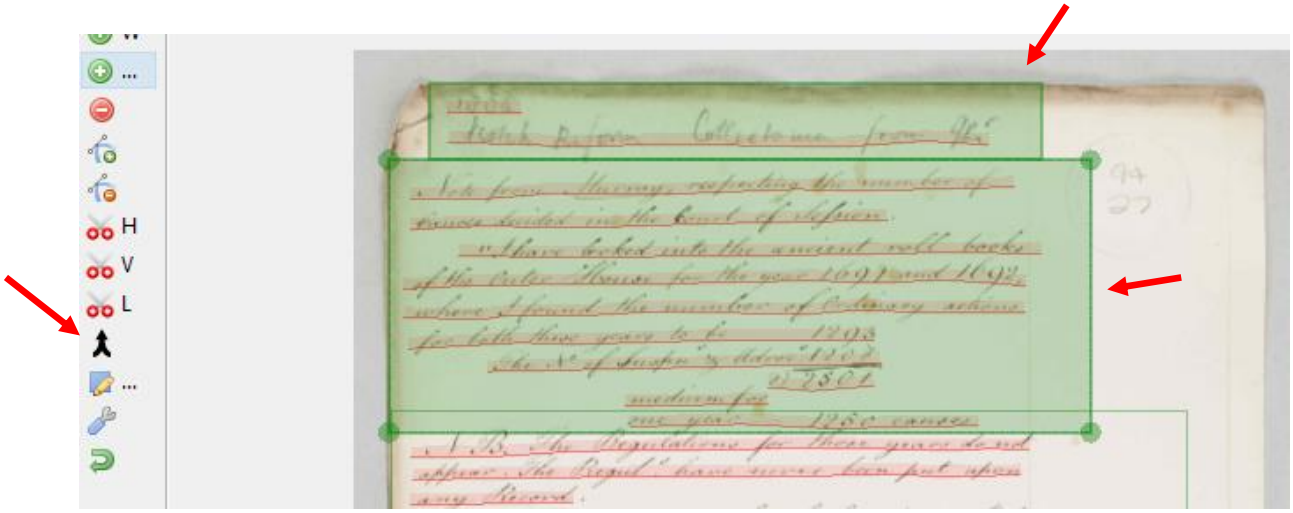


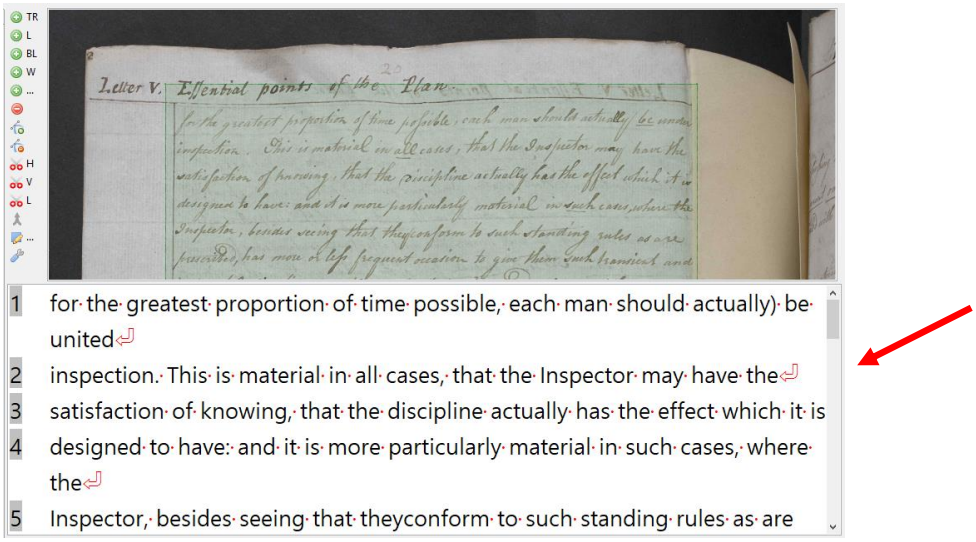
Figure 8 Merge two text regions

Correct baselines

- Of course it is also possible to correct baselines in your document.
- As with the text regions, click on a baseline and you can drag the parts of the line, split a line into two or merge two lines together.
- You can also delete a baseline and draw a new one from scratch. Click the "+BL" button in the Canvas menu. Click once to start drawing your baseline and double-click to finish your line.
- Note: Baselines are most important for "Handwritten Text Recognition"; line regions do not need to be corrected.

Transcribe text

- Select the "Transcription" viewing from the Main menu.
- You will see the text editor field below the image: **For each line/baseline in the image you will find a corresponding line in the text editor**. The image and the text are connected in this way.



1 for the greatest proportion of time possible, each man should actually) be united
2 inspection. This is material in all cases, that the Inspector may have the
3 satisfaction of knowing, that the discipline actually has the effect which it is
4 designed to have: and it is more particularly material in such cases, where the
5 Inspector, besides seeing that theyconform to such standing rules as are

Figure 9 Transcribe your document

- Transcribe the text according to the language of your source document. Use the characters of your keyboard.

- **You can have more than one person working on a document but they should not work on the same page simultaneously.** You can let other Transkribus users see your documents by clicking the "User Manager" button in the "Server" tab.
- **Note: "Handwritten Text Recognition" can work on both handwritten and printed documents.** Simply upload, segment and transcribe your document in the same way as described above.

Train a model

- In order to train a model, we recommend that you transcribe at least 15,000 words (around 75 pages).
- After that just drop us a short email (email@transkribus.eu) and we will get back to you about the training of your model. You can also check the corresponding How To Paper on the Transkribus Wiki Site.

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 H2020 e-Infrastructure Project READ (Recognition and Enrichment of Archival Documents) which received funding from the European Commission under grant agreement No 674943.