READ
Recognition and Enrichment of Archival Documents

How To Train A Handwritten Text Recognition Model In Transkribus

Version v1.8.0
Last update of this guide: 24/10/2019

This guide explains how to use Transkribus to train a Handwritten Text Recognition (HTR+) model to recognise your documents. After training the model, will help you to automatically transcribe and search your collection.

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

- The Transkribus platform allows users to train a Handwritten Text Recognition (HTR+) model to automatically process a collection of documents. The model needs to be trained to recognise a certain style of writing by being shown images of documents and their accurate transcriptions.
- For the training of a model between 5,000 and 15,000 words (around 25-75 pages) of transcribed material are required. If you are working with printed rather than handwritten text, a smaller amount of training data is usually required.
- The model training function is not automatically included in the standard Transkribus platform. When you are ready to train a model, contact the Transkribus team (email@transkribus.eu) and they will give you access to the feature.

Preparation

- We recommend that you start the training process with between 5,000 and 15,000 words of transcribed material, depending on if it is printed or handwritten text.
- The neural networks in HTR+ learn quickly and the more training data they have, the better the results will be.
- You can create training data for HTR+ in Transkribus by uploading images and transcribing text. For full instructions, see How To Transcribe Documents with Transkribus - Introduction.
- If you already have existing transcripts, you can also use these to train your model. For more information see How To Use Existing Transcriptions to train a HTR model.

Training

- The main options for the training of a model can be found in the **“Tools” tab** in the **“Text Recognition”** section.
- As **“Method”**, “HTR (CITlab)” is the most effective option to choose.
- By clicking the “Models” button you can see which models are available and which documents they were trained on.
- With the **“Train”** button you will arrive at the options for the training of models.

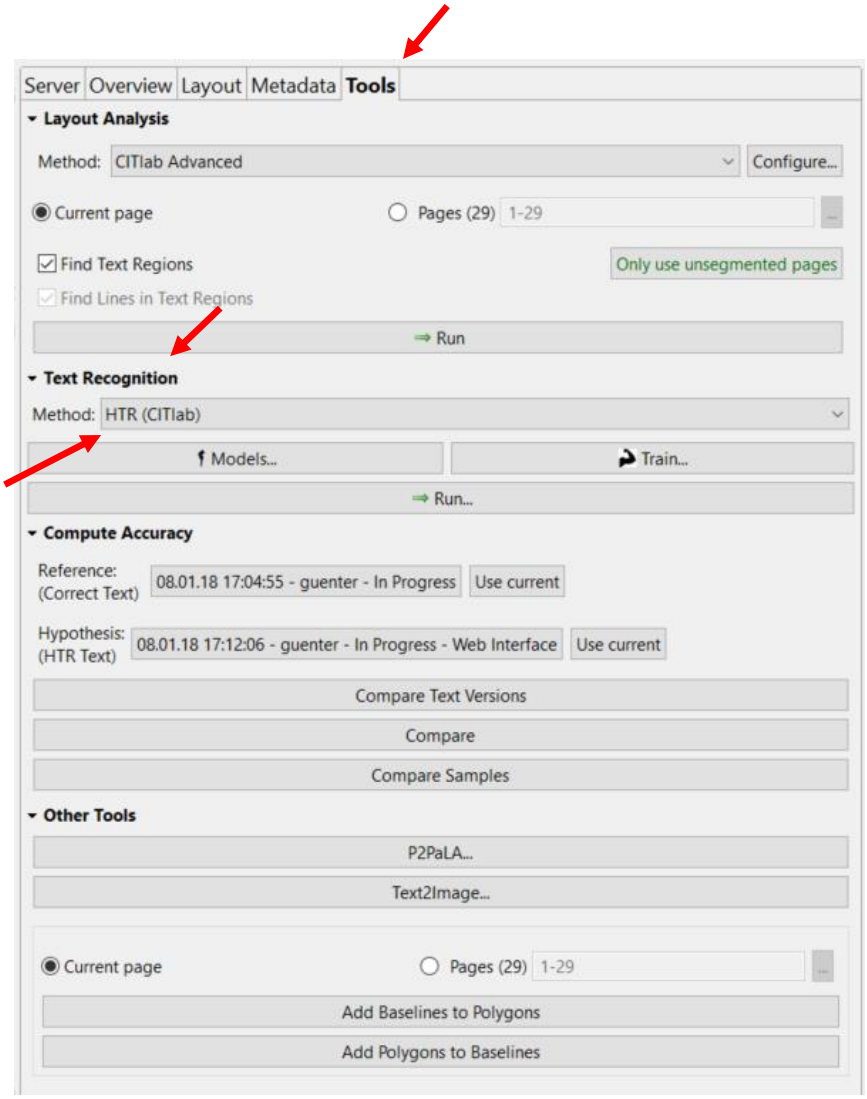

Figure 1 Where to find the tools for the training

Setting up HTR+ Training

- To get to the "HTR+ Training" window, click the "Train" button in the "Tools" tab.

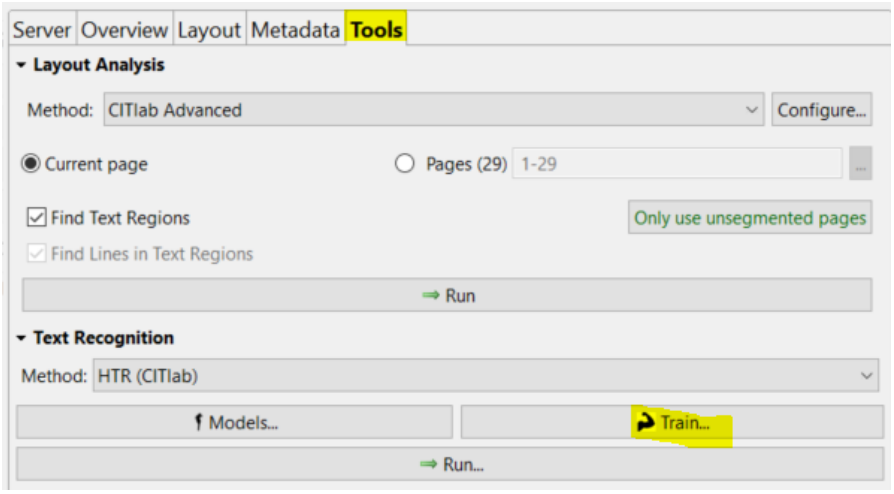

Figure 2 How to open the "HTR Training" window.

- The following window will open up:

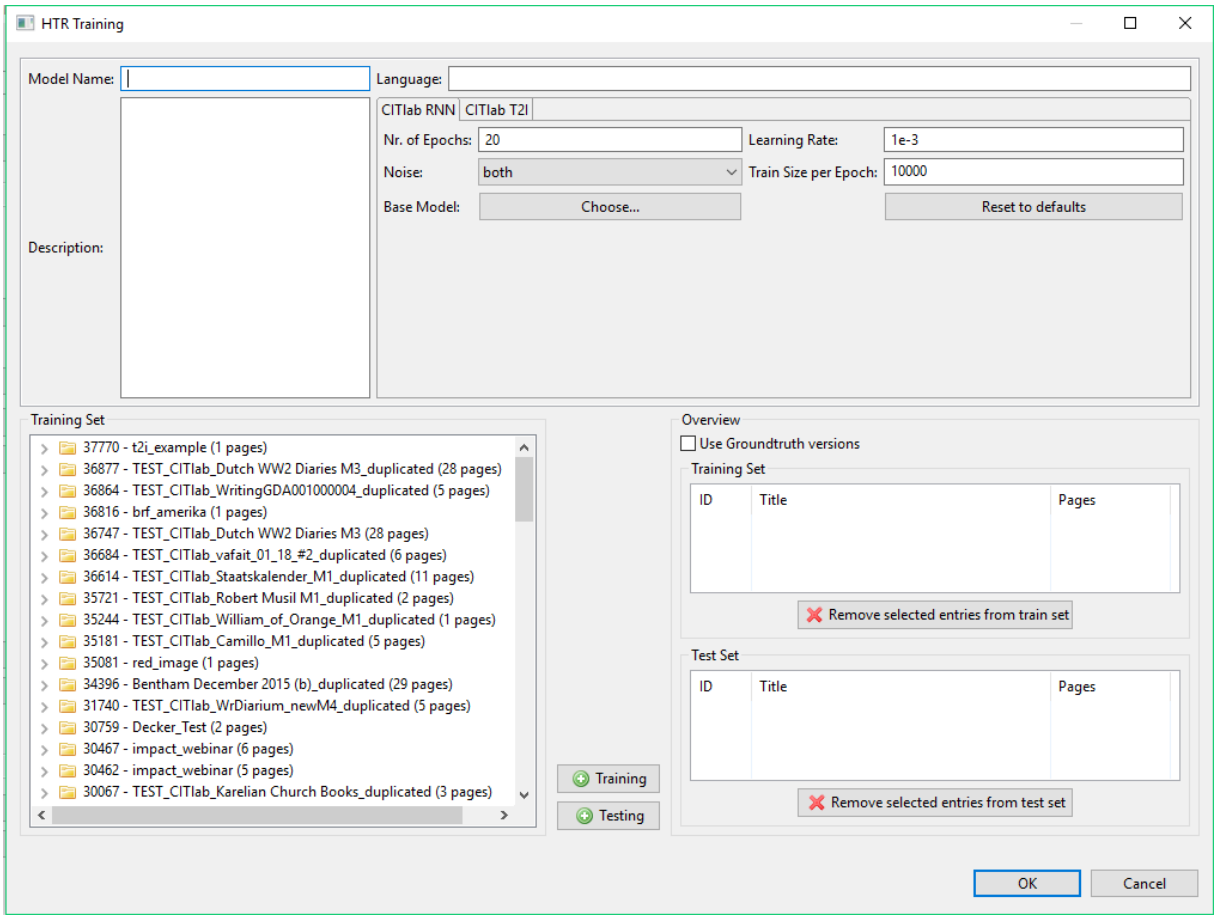

Figure 3 "HTR Training" window

- In the upper section you will need to add details about your model.

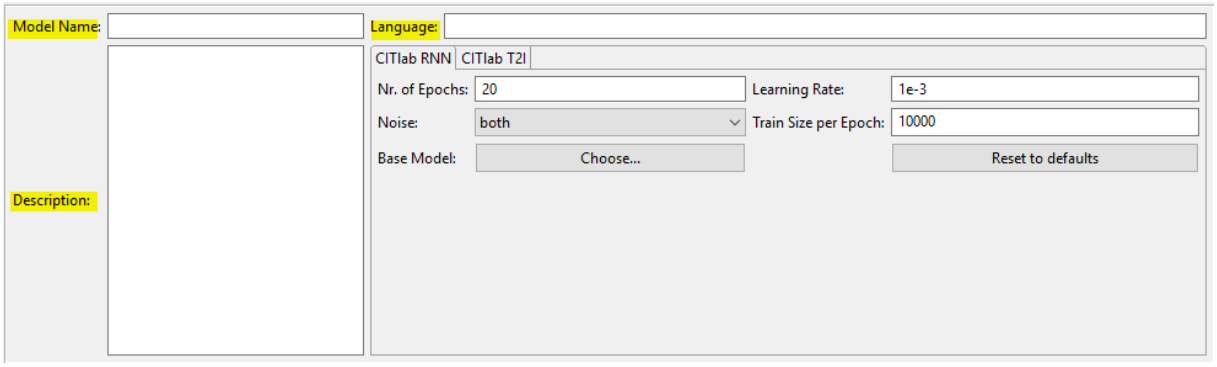

Figure 4 Adding details about the model

- Please add
  - Model Name (chosen by you)
  - Language (of your documents)
  - Description (of your documents and the pages selected as training and test data)
- Note: "Nr. of Epochs" refers to the number of times that the training data is evaluated. If you increase the number of epochs, the training process will take longer.

Base Model

- It is possible to add a base model to your training. If you choose this option, the information the base model contains will be integrated to the new model. To have a benefit the base

model needs to be similar to the writing it should recognise afterwards. With the help of a base model it is possible to speed up the training process. An improvement of quality is not guaranteed, it has to be tested in the individual case.

- One big benefit of working with base models is, that they can make it possible to start with a smaller amount of training pages, which means that the transcription workload would be reduced.
- To use a base model, you simply need to choose the desired one with the “Choose…” button next to “Base Model:”.

Training Set

- Next, you need to select the pages that you would like to be included in your set of training data.
- To add all the pages of your document to the Training Set, click on the folder and click “+Training”.
- To add a specific sequence of pages from your document to the Training set, double-click on the folder, click on the first page you wish to include, hold down the “Shift” key on your keyboard and then click the last page. Then click “+Training”.
- To add individual pages from your document to the Training Set, double-click on the folder, hold down the “CTRL” key on your keyboard and select the pages you would like to use as training data. Then click “+Training”.
- The pages you have selected will appear in the “Training Set” space.

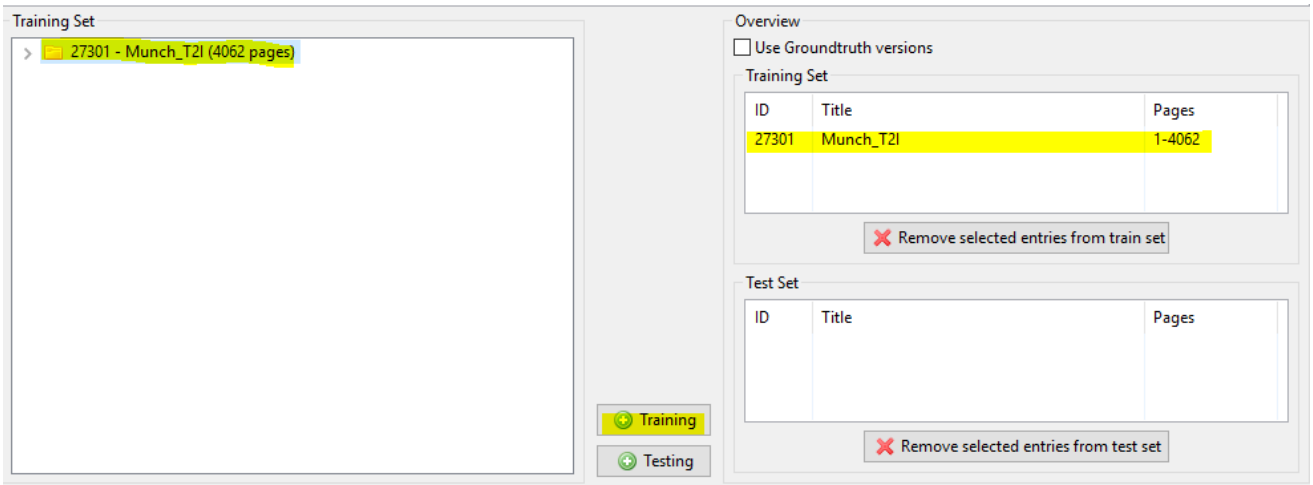

Figure 5 Adding all the pages for training

Test Set

- During the training process, a Test Set of pages is set aside and is not used to train the HTR. These test pages can then be used to assess the accuracy of your model.
- We recommend that you select at least one test page for every 50-100 pages of your Training Set.
- The pages in your Test Set should be representative of the documents in your collection.
- The more pages there are in your Test Set, the longer the HTR training will take.
- To add pages to the Test Set, follow the same process as above but click the “+Testing” button.

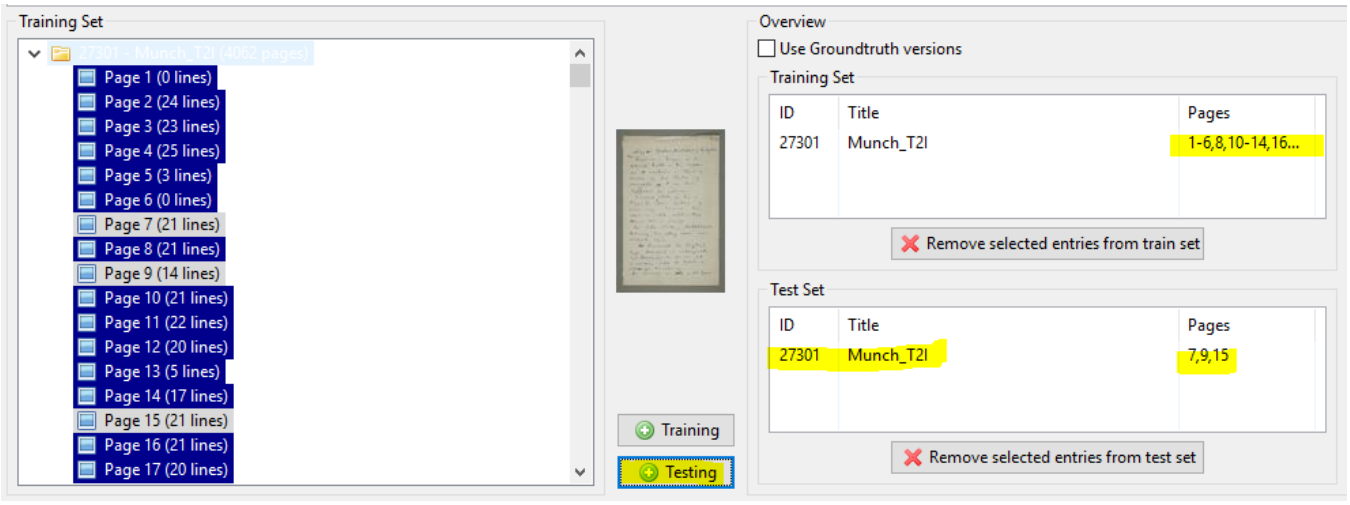

Figure 6 Adding pages to the Test set

- To remove pages from the "Training Set" or "Test Set", click on the page and then click the red cross button.

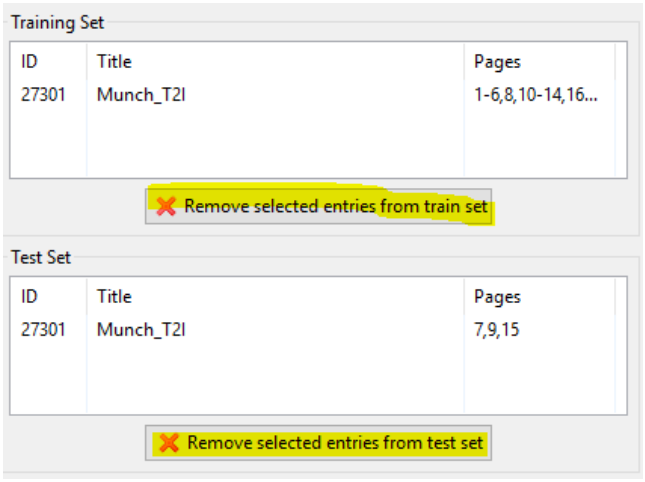

Figure 7 Removing pages

- You can make a note of the pages used in your test set in the model description box.
- Start the training by clicking the "OK" button.

Checking progress

- You can follow the progress of the training by clicking the "Jobs" button in the "Server" tab.

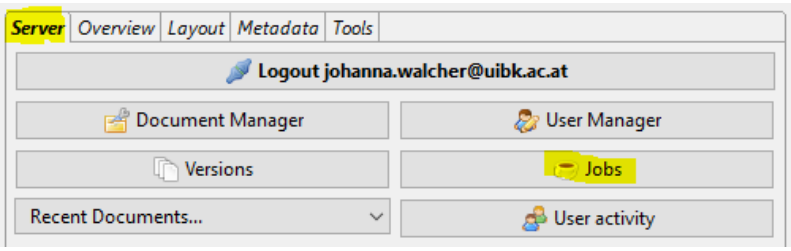

Figure 8 Check the progress of the training with the "Jobs" button

- The completion of every epoch will be shown in the "Jobs on server" window, as well as the completion of the training process.

- Training a HTR+ model will take at least a couple of days. You can perform other jobs in Transkribus or close the platform during the training process.

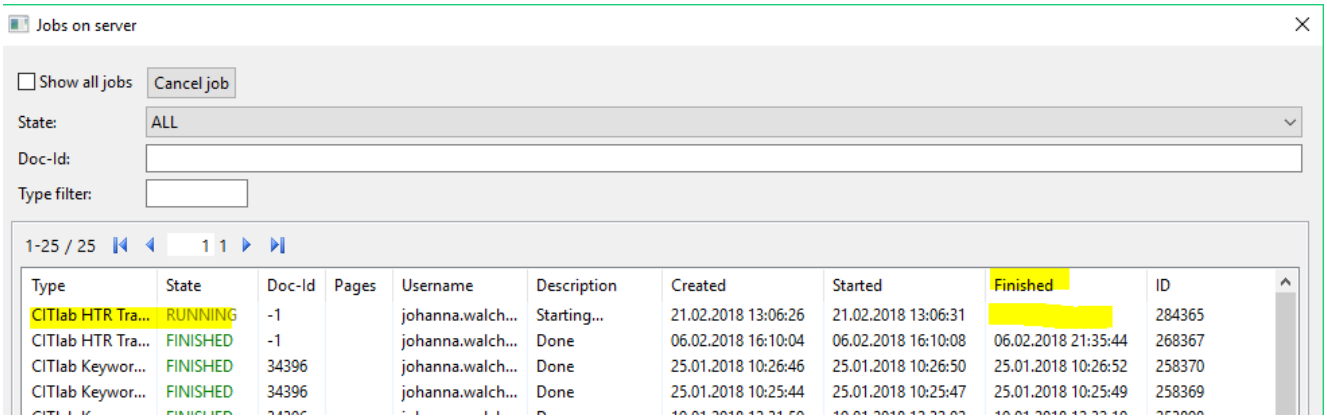

Figure 9 "Jobs on server" overview

After the training

- After the training of your model is finished it will be available in your collection.
- In order to access it click the "Models" button in the "Tools" tab.

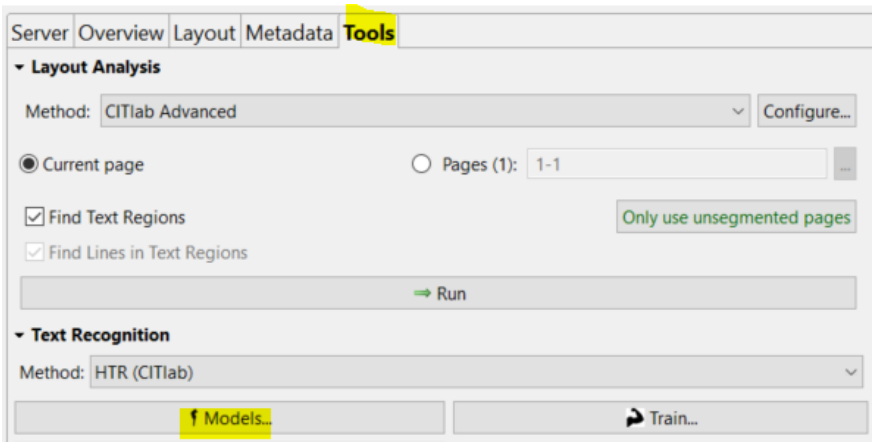

Figure 10 Opening the "Choose a model" window

- The following window will open up:

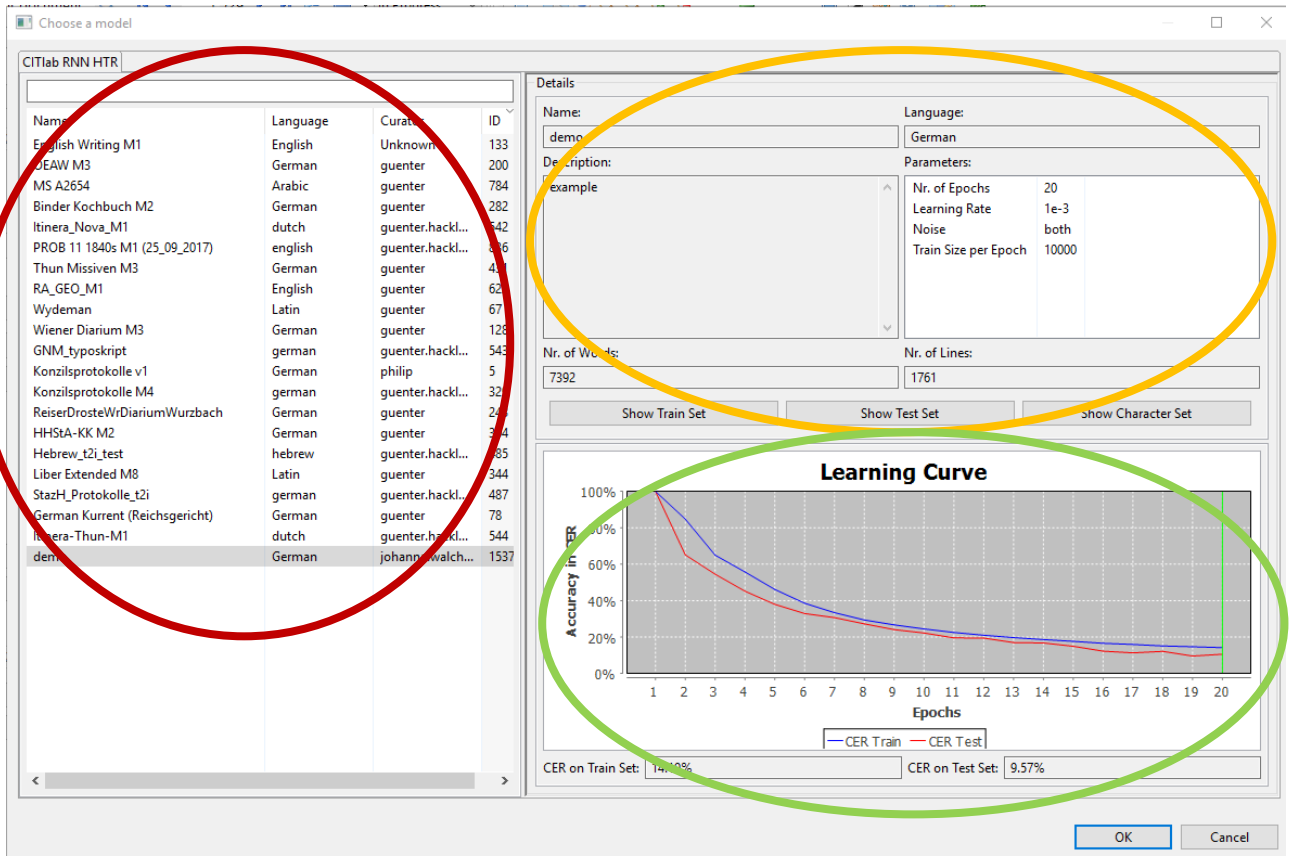

Figure 11 "Choose a model" window

- On the left side of the window you can see an overview of the available models.
- On the top right side of the window the details of the model are shown.
- On the bottom right you can see the learning curve of your model. More information about these statistics can be found below.

Statistics

- The "Learning Curve" graph signifies the accuracy of your model

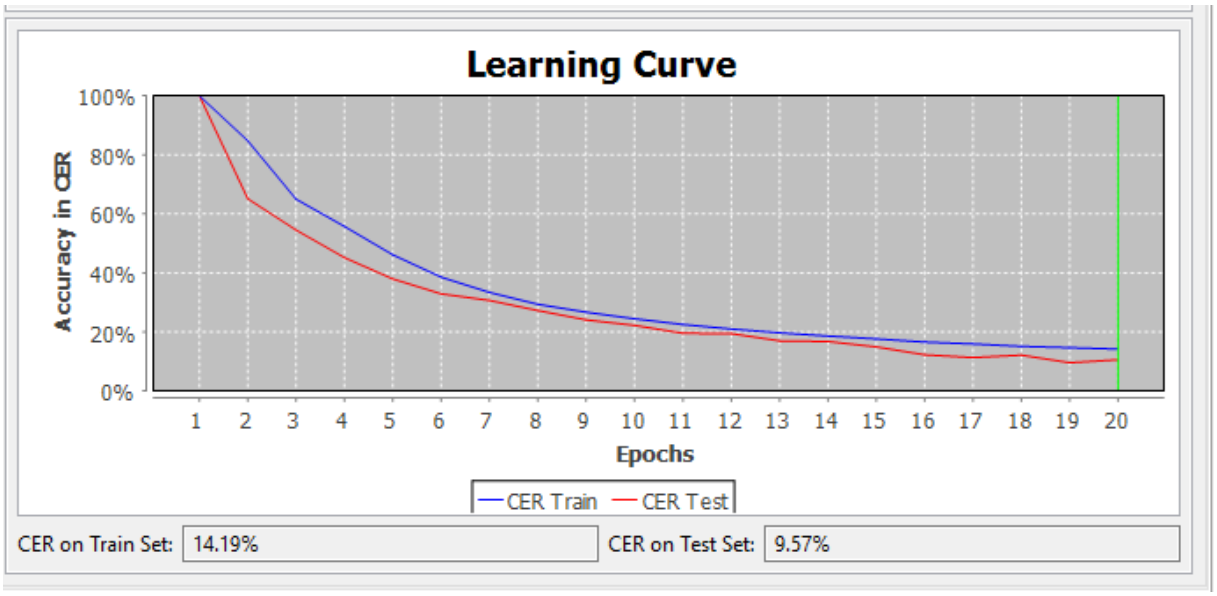

Figure 12 "Learning Curve" of your model

- As you can see in Figure 12 the y-axis is defined as “Accuracy in CER”
- “CER” stands for **Character Error Rate**, i.e. the percentage of characters that have been transcribed incorrectly by HTR+.
- “**Accuracy in CER**” is indicated as percentage on the y-axis. The curve will always start at 100% and will go down as the training progresses and the model improves.

- The x-axis is defined as “**Epochs**”.
- During the training process Transkribus will make an evaluation after every epoch. In Figure 12 the “Training Set” was divided into 20 epochs.
- When you train a model you can indicate how many “epochs” the “Training Set” should be divided into. The more epochs there are, the longer the training will take.

- The **graph** shows two lines, one in blue and one in red.
- The **blue line** represents the progress of the training.
- The **red line** represents the progress of evaluations on the Test Set.
- First the program trains itself on the **Training Set**, then it will test itself on pages in the **Test set**.
- Underneath the graph, two percentage values are shown relating to the CER for the Training Set and the Test Set.
- In Figure 12, the model performs with a 14.19% CER on the Training Set and 9.57% on the Test Set.
- The value for the Test Set is the most significant as it shows how the HTR+ performs on pages that it has not been trained on.
- Results with a CER of 10% or below can be seen as very efficient for automated transcription.
- Results with a CER of 20-30% are sufficient to work with powerful Keyword Spotting technology. For more details, see our How To Transcribe – Keyword Spotting guide.

Generating HTR transcripts

- Now that you have your model, you can use it to automatically generate transcripts of the documents in your collection.
- First, **upload** your documents to Transkribus.
- Second, **segment** your documents into text regions, lines and baselines.
- For more information on **uploading** and **segmentation**, please consult How To Transcribe Documents with Transkribus – Introduction.
- To access your model, click on the “Tools” tab and go to the “Text Recognition” section.
- Click “Run”, then click “Configure”. Choose your HTR model from the list on the left-hand side of the screen and click OK.
- Select whether you wish to generate a HTR transcript of one page or several pages.
- Press “Run” to start the text recognition process.
- Once the recognition is finished, the automated transcription will appear in the text editor field.

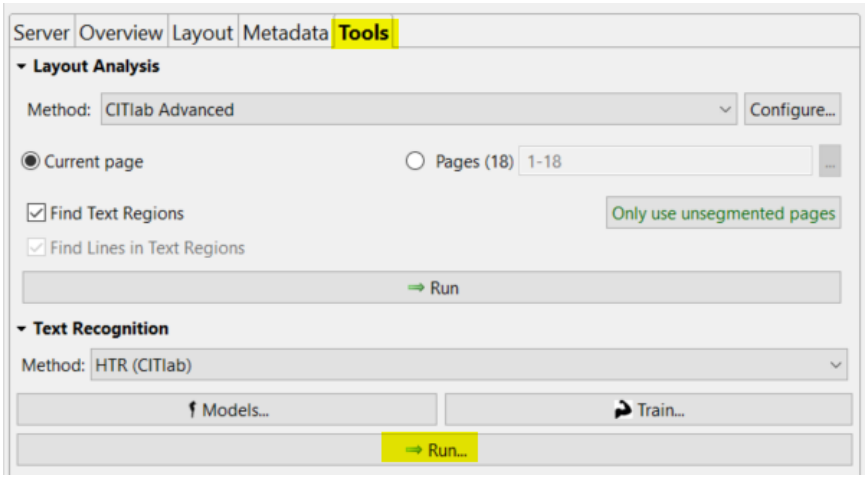

Figure 13 Run model

Share a model

- You can share your HTR model with other collections in Transkribus, whether they are owned by you or by other users.
- If you want to share your model with another collection, you must have access to that collection.
- Right click on the name of your model (on the left side of the "Choose a model" window).

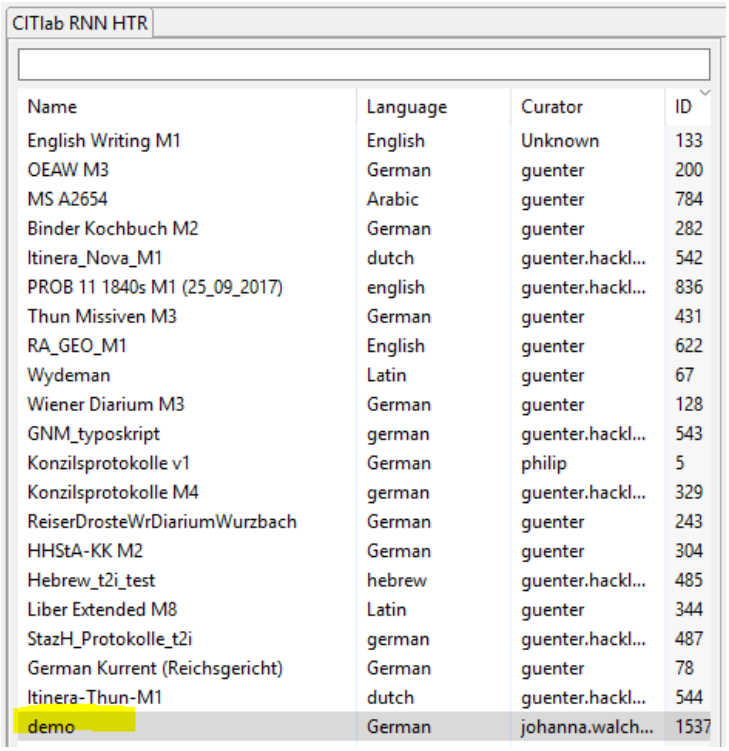

Figure 16 Share a model by right-clicking the name of your model

- Then select "Share model…"
- The "Choose a collection via double click" window will open up.
- In the next window click the collection you would like to share the model and press "OK".
- In this window, you can also create a new collection for the model with the "Create" button.
- Click "OK" to confirm.

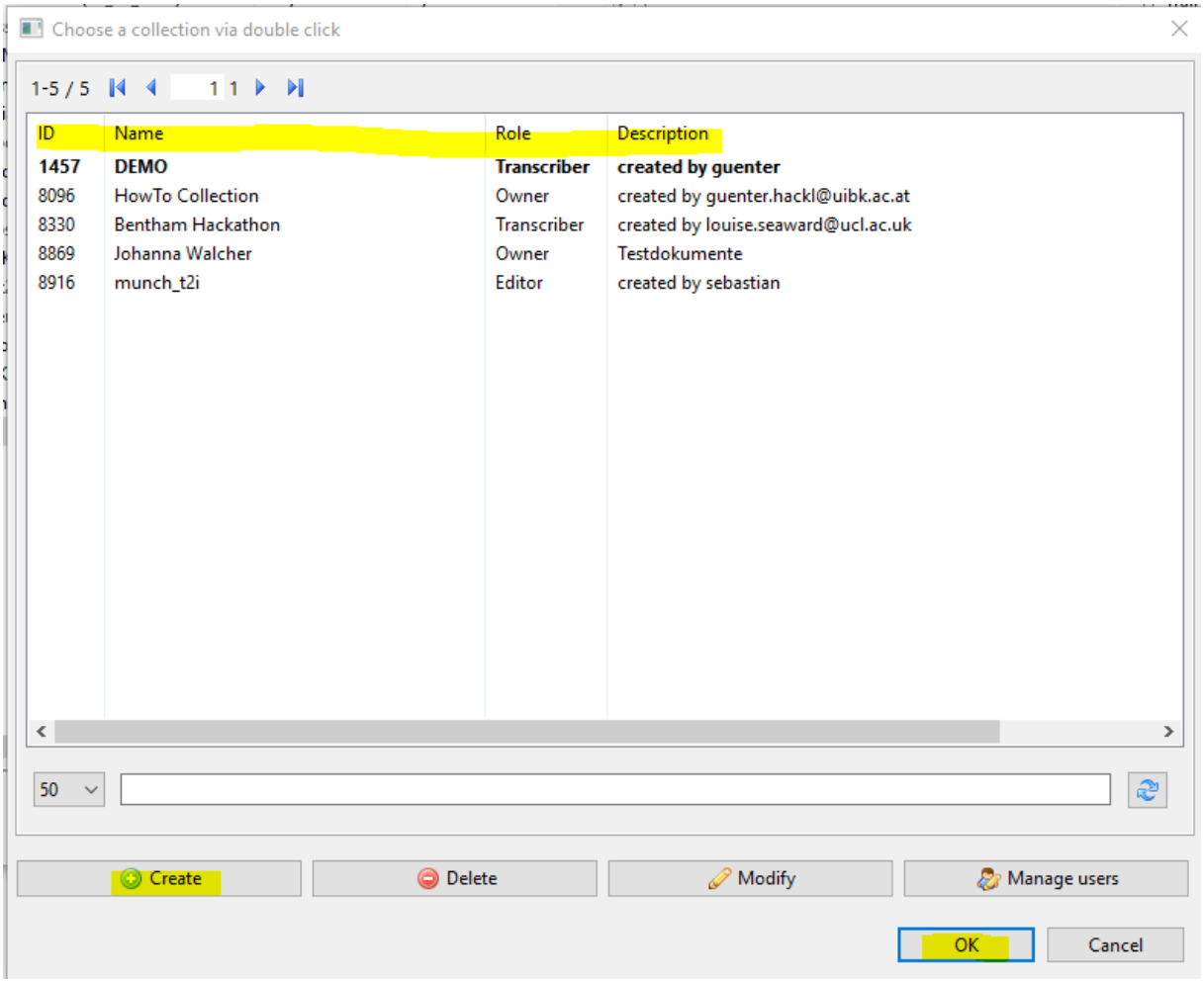

Figure 17 How to share your model

- Once you have chosen the collection, click “OK” once more and the model will now be shared.

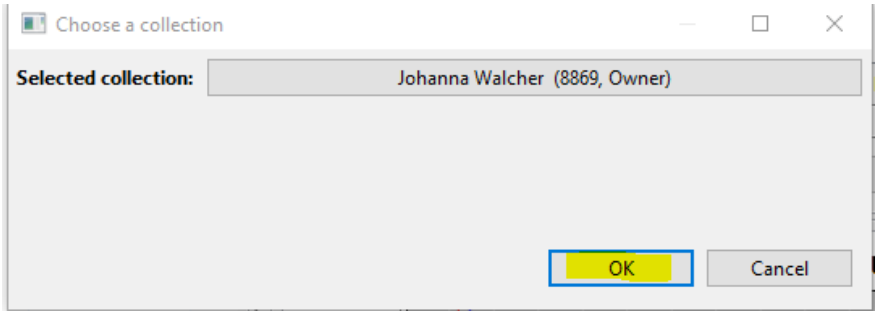

Figure 18 Confirm the sharing of your model

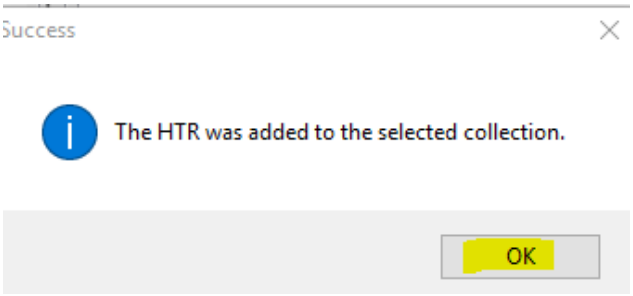

Figure 19 Model had been shared

Your output

- As soon as the training is finished, you can try out your model on any other historical document with similar writing.
- You can share your model with other people who can benefit from it too.
- You can repeat the training process with more data in order to generate more efficient results.
- You can measure the accuracy of your model with the “Compute Accuracy” function.
- The results of the HTR will depend on how similar and how clear the writing in the historical document is.
- The Transkribus team is working on an algorithm which will make it possible to automatically transcribe any kind of document, without the need to prepare training data. The technology is learning from all training data processed in Transkribus.
- So the more data we work with, the more efficient the technology will become. Train your own model and be part of it! 😊

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.