READ
Recognition and Enrichment of Archival Documents

How To Prepare Test Projects with Transkribus - for Archives and Libraries

This is a short introduction to the basic steps for setting up a test or pilot project with Transkribus. When you work with the platform, you will gain access to data on the state-of-the-art in Handwritten Text Recognition. You will also be able to apply this technology to a set of documents.

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

- As an archive or library you may be responsible for a (large) digital collection of handwritten documents, e.g. several thousands, tens-of-thousands or even millions of documents.
- You want to make these documents available and allow users to search your collection in a previously unknown way using full-text, keyword spotting or named entity recognition
- You are interested in using Handwritten Text Recognition (HTR) technology to automatically generate transcripts of the documents in your collection
- You would like to know how HTR technology works and how accurate it can be

Designing a test project

- With Transkribus it is really simple to start a test project which will enable you
    - to train the HTR engine to recognise a specific collection
    - to evaluate the accuracy of the text recognition process in a scientific way
    - to extrapolate your results to the entirety of your collection and
    - to estimate the time and resources needed to process your complete collection.
- The test project can be done either by you or your staff or we can assist you and manage it for you (details below).
- You can start the test project right now. The tools and workflows are already available and can be started via the Transkribus expert interface.

Training an HTR engine

- HTR engines are based on machine learning algorithms, more specifically on supervised machine learning. This means that the HTR engine needs to be shown “correct examples” of documents that have been transcribed in order to understand the patterns forming characters and words.
- In general one can say that the more training data there is, the more accurate the results. This is especially true for collections which contain a large number of different writers or writing styles.
- Research groups within the READ project are working on a unified model which will integrate all training data. If similar documents to yours have already been used as training data, it will be easier to produce HTR for your collection.

Selecting the dataset

- In order to run the test project you will need a reliable dataset. This should be a representative sample of the documents contained in your collection. If you use a reliable data set for your test project, the results should be able to be extrapolated to the rest of your collection.
- The data set will serve as training material for the HTR engine and will also be used to evaluate the results of the HTR.
- We recommend that you select documents randomly from your collection. The most objective way is to select documents automatically using a database or by applying a simple criteria such as selecting every 10^{th}/20^{th}/50^{th} document.

How many pages to select?

- As a rule of thumb, 20,000 words (around 100 pages) can be sufficient for training if you are dealing with a simple collection of documents, e.g. a diary, or letters written by one person.
- If you have multi-writer documents and/or collections spanning several decades or centuries, we recommend that you select a dataset of several hundred pages.
- Nevertheless first tests can always be done on a small data set and then the amount of training data can be increased according to the results achieved by the HTR engine.

The dataset becomes ground truth

- The dataset needs to be prepared in a specific way so that the HTR engine is able to use it as a "learning resource".
- This data is known as "ground truth" or "reference data" since it forms the basis for all other operations.
- There are two main principles which have to be followed when creating "ground truth" data:
    - Segmentation
    - The lines of the transcript must be connected with the lines of the image, so that the computer is able to match each line of the transcript with its corresponding line in the image. To achieve this, each image must be segmented into text regions, lines and baselines. More details can be found in: How to Transcribe Documents with Transkribus – Introduction.
    - Transcription
      The dataset needs to contain the correctly transcribed text of the document. The text should be as close to the actual appearance of the document as possible, e.g. every letter in the document should be represented by its corresponding character in the transcript.
- For modern documents, e.g. from the 18th century onwards, the transcription of documents is usually straightforward. In documents from earlier periods, issues like unusual characters and scribal abbreviations can raise some challenges. But do not worry, all these issues can be handled with the tagging system included in Transkribus.

Transcribing the dataset

- Once you have selected the pages for your dataset, they can be uploaded to Transkribus and the transcription process can start. Note: all documents in Transkribus are private; no other users have access to your documents.
- There are two ways to carry out the transcription on Transkribus:
    - Option 1: Do it yourself
      Either you, your colleagues or staff will transcribe the text in Transkribus. In this case, you need to learn to use Transkribus which may take 2-3 hours of "learning by doing". The detailed instructions for segmentation and transcription can be found in: How To Transcribe Documents with Transkribus – Introduction.
    - Option 2: We do it for you
      If you have existing transcriptions, we have a new Text2Image matching tool that can match these to digitised images automatically. Alternatively, students and external service providers will produce the transcription. They have experience of working with Transkribus and old European languages. The price will depend on the required level of accuracy and the difficulty of the handwriting.

HTR training

- Once the dataset is completely transcribed and has become a valid "ground truth", the process of training the HTR engine can begin.
- This is an offline process carried out within the Transkribus platform. It will very likely take a few weeks until a new or updated HTR model is available.
- Once the training process is completed you will receive a notification and will then be able to use the HTR model to automatically transcribe and search the rest of the documents in your collection.

Evaluating the results

- Your data set will not only be used for training purposes. A small part of it will be set apart and used as a test set. The images of the test set will not be used to train the HTR engine.
- The test set will therefore allow you to evaluate the accuracy of results of the HTR engine.
- Transkribus offers you a mechanism to carry out this evaluation directly on your documents. For this purpose a module has been implemented in Transkribus which computes the Character Error Rate and the Word Error Rate, both approved metrics in computer science.
    - Note: Once a HTR model has been trained and is available via Transkribus you may apply it to any page – including those which were not part of the initial data set. You can measure the accuracy of the HTR for a specific page or pages as well.
- Current results from computer science show that Character Error Rates below 10% and Word Error Rates below 20% represent the state-of-the-art in HTR technology. In lab conditions even better results can be achieved.

What you will get from the HTR process

- The HTR engine will produce an automatic transcript of your documents, and more besides.
- The HTR engine will also produce confidence matrices on character and/or word level. This is a way of storing the internal options that the HTR engine was considering.
- Based on such confidences, two additional functions can be applied to your documents:
    - Keyword Spotting
      This is a technique of searching for words which are not part of the "first best" transcription generated by the HTR engine. Keyword Spotting searches for words throughout the internal options of the HTR engine. This increases the chance of finding the correct word, even if it has not been assigned the highest confidence rating. E.g. instead of missing every fifth word (Word Error Rate 20%), you will probably be able to find 95% of all words by using Keyword Spotting. Keyword Spotting is now available within the Transkribus expert client.
    - Computer Assisted Transcription
      Confidence matrices are also useful when a user is correcting a transcribed page. Alternative words can be shown to the user, or the interface can offer likely suggestions of words based on recent input from the user. A demo version of Computer Assisted Transcription is available on the website of the tranScriptorium project.

What about structural data?

- Many archival documents are organized in a structured way, i.e. in tables or forms with repeating elements. Of course such additional information can be used to increase the value of the automated transcription process.
- Good results from the HTR engine are an important prerequisite for any kind of structural enrichment. Moreover structural information is less normalized than writing styles and has to be treated manually, e.g. by applying rule based systems. This requires a lot of expert knowledge as well as the involvement of programming staff.
- We recommend that you carry out initial tests focusing just on the reliability of the HTR and afterwards, take the opportunity to enrich the data with structural information.

Next steps

- If you are interested in carrying out a test project with (parts of) your collection, we recommend that you contact us beforehand (email@transkribus.eu) so we can clarify the main cornerstones of your project.
- You may also consider becoming part of the READ project via a Memorandum of Understanding (MOU). This will give you the chance to look behind the scenes and receive information first hand. Take a look at the READ project website to see a list of libraries and archives who have already signed a MOU.

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.