

Transkribus: Transkription und Annotation - historische Manuskripte, alte Drucke – oder Social Media

Was ist Transkribus?

Transkribus ist eine Plattform zur Erkennung, Digitalisierung und Transkription von handschriftlichen und gedruckten historischen Dokumenten. Transkribus bietet darüber hinaus eine computergestützte Handschriftenerkennung, Bilderkennung und Strukturerkennung an. Transkribus wird durch die 2019 gegründete READ COOP SCE erhalten und weiterentwickelt. Die Homepage der READ COOP [Transkribus | Handwritten Text Recognition | READ COOP](#) bietet ausführliche Guides und umfangreiche Informationen zur Plattform. Der hier präsentierte Kurzleitfaden bietet einen Überblick für den Einstieg in die Arbeit mit Transkribus. Die hier angeführten Informationen sind vertiefend und umfangreich in den Guides dargestellt. Diese werden zur vertiefenden Einarbeitung in das Programm empfohlen.

Transkribus ist auch sehr gut geeignet, um jüngere Bilddateien wie Screenshots aus Social Media systematisch zu bearbeiten und Tags hinzuzufügen.

Erstellung eines Accounts und Installation

Zunächst legen sich Nutzer*innen einen Account auf der Homepage der READ COOP an: [Transkribus | Handwritten Text Recognition | READ COOP](#). Anschließend kann das Programm heruntergeladen werden. Es ist mit Windows, Mac und Linux kompatibel und erfordert die neueste Version von Java.

Vorbereitung und Import von Dokumenten

Es können zum Beispiel Scans (pdf) des Originaldokuments für die Arbeit mit dem Programm verwendet werden. Das Dokument kann in das Programm importiert werden (Hauptmenü Import document(s)) und zu einer eigenen Bibliothek hinzugefügt werden. Die Dokumente in der eigenen Bibliothek sind geschützt und werden nicht geteilt.

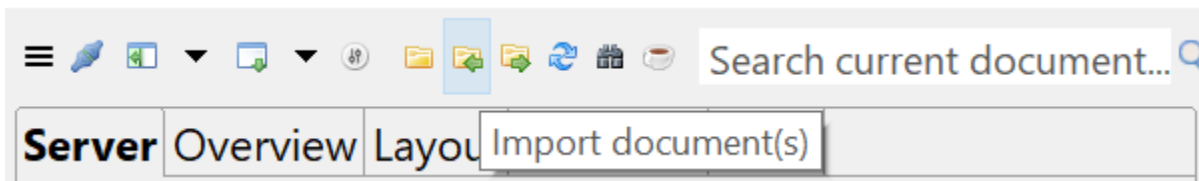


Abb 1.: Dokument(e) importieren

Segmentierung

Bevor mit der Transkription begonnen werden kann, wird das Dokument segmentiert. Die Einteilung in Textblöcke, Grundlinien und Zeilen erfolgt automatisch durch das Programm.

Der Reiter "Tools" wird aufgerufen. Layout Analysis Current page Find text regions Run

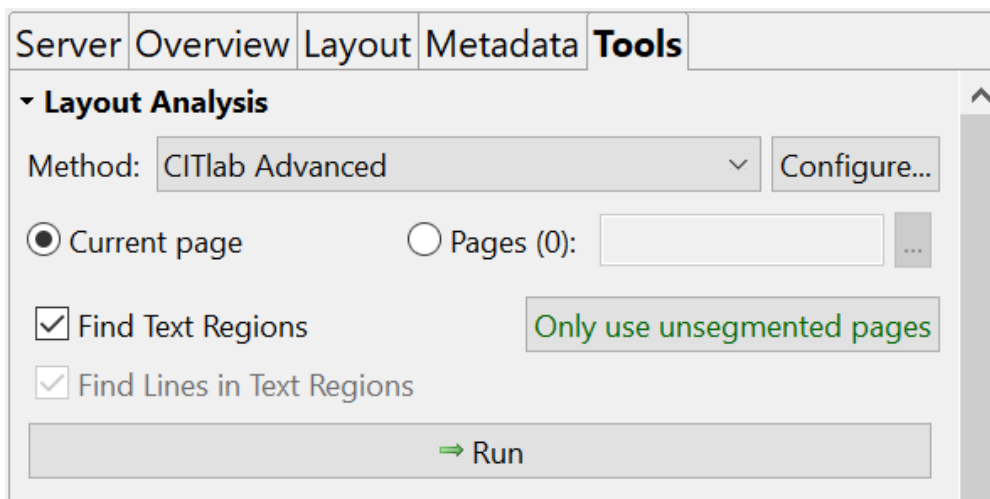


Abb. 2: Segmentieren

Transkribieren

Mit der manuellen Transkription kann erst nach der Segmentierung begonnen werden. Auch die Handschriftenerkennung (HTR) kann erst trainiert werden, wenn die automatische Segmentierung des Dokuments erfolgt ist.

Ist das Dokument segmentiert, kann der Text Zeile für Zeile transkribiert werden.

Die einzelnen Komponenten (Textblöcke, Grundlinien und Zeilen) können bei Bedarf manuell verändert und angepasst werden.



Abb. 3: "Werkzeuge" zur manuellen Nachbearbeitung von Textblöcken, Grundlinien und Zeilen

Tags und Metadaten

Das Dokument kann zusätzlich mit Informationen und textstilistischen Merkmalen angereichert werden.

Der Reiter "Metadata" wird aufgerufen Tags All Tags

Unter dem Menüpunkt "All Tags" kann aus einer Auswahl vordefinierter Tags gewählt werden. Um den Text mit einem Tag zu versehen, wird der gewünschte Bereich markiert und durch Anklicken von "+" mit dem jeweiligen Tag versehen.

Bei der Arbeit mit privaten Dokumenten kann die Verwendung des *Blackening Tag* hilfreich sein, welcher die Anonymisierung sensibler Daten im Text ermöglicht.

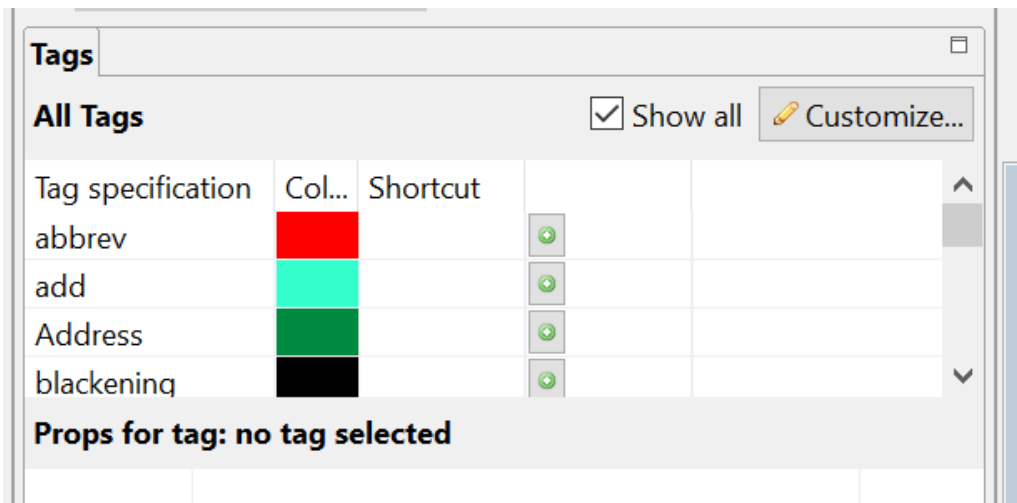


Abb. 4: Text mit Tags versehen

Individuelle Tags erstellen

Neben den bereits vordefinierten Tags können auch individuelle Tags definiert werden.

Der Reiter "Metadata" wird geöffnet Textual Tags All Tags Customize Create new tag

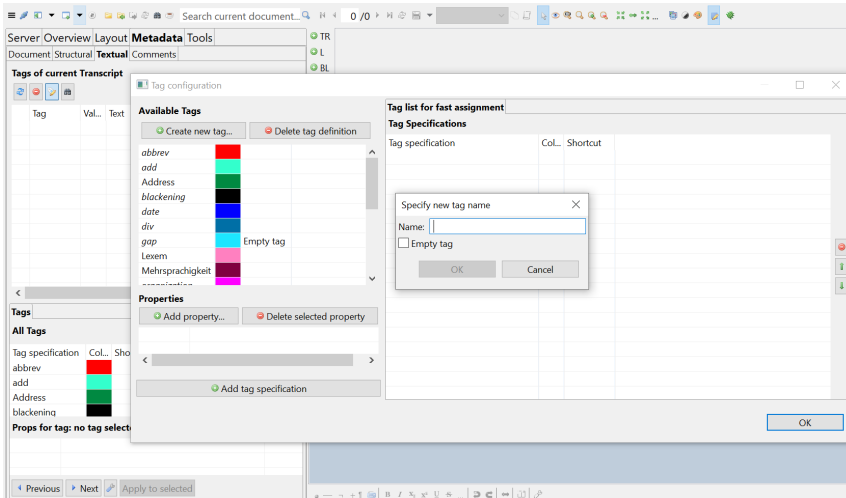


Abb. 5: Individuelle Tags erstellen

Erstellen eines Metadatenabschnittes

Metadaten zum Dokument können unter dem Reiter "Metadata" Document erfasst werden.

Die Metadaten zum Dokument werden dem fertigen Transkript nach Export vorangestellt.

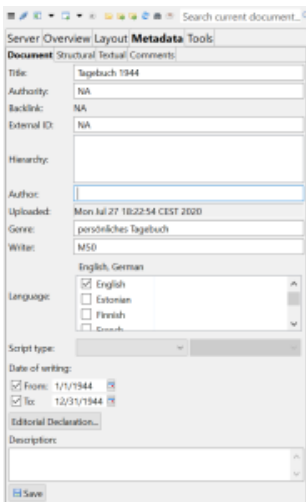


Abb. 6: Metadaten einpflegen

Speichern und Export

Das Dokument kann in der eigenen Sammlung gespeichert werden. Eine Bearbeitung in mehreren Etappen ist somit möglich. Sobald das Transkript fertiggestellt ist, kann es exportiert werden.

Es wird im Hauptmenü "Export documents" angewählt.

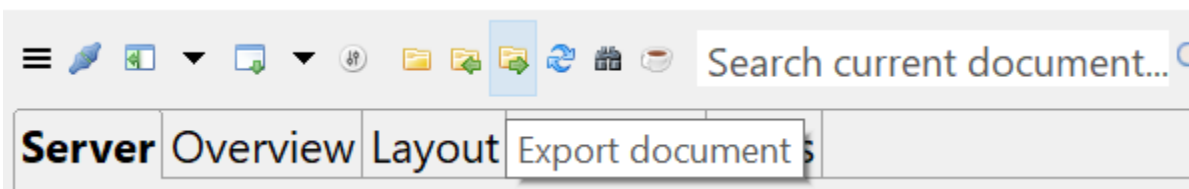


Abb. 7: Dokument exportieren

Das zu exportierende Dokument kann in verschiedenen Formaten ausgegeben werden.

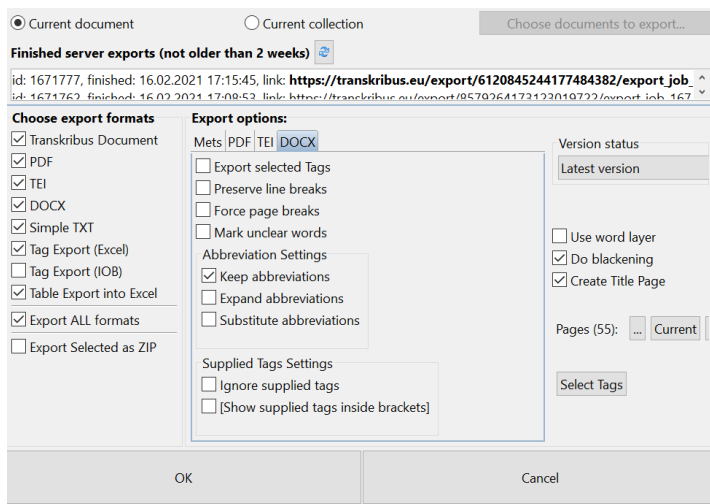


Abb.8: Auswahl an Formaten (Export)

Hinweis: Wird ein Haken im Feld "Do blackening" gesetzt, werden mit dem *Blackening Tag* versehene Textpassagen und Wörter im exportierten Dokument anonymisiert.

Wird ein Haken im Feld "Create Title Page" gesetzt, werden die im Metadatenabschnitt eingepflegten Daten exportiert und dem Transkript in der exportierten Datei vorangestellt.

Weitere *Tags* können in bestimmten Exportformaten, z.B. DOCX oder TEI exportiert werden.

Das fertige Transkript wird den Nutzer*innen nach dem Export per E-Mail zugestellt.