

DigEdTnT - Webinarreihe

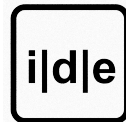
Tools & Transitions II

Transkribus → FairCopy

<https://digedtnt.github.io>

03.10.2023

We work for
tomorrow



CLARIAH-AT



ZENTRUM FÜR
INFORMATIONSMODELLIERUNG
AUSTRIAN CENTRE FOR
DIGITAL HUMANITIES





Transkribus®



ediarum



ba[sic?]



Bild & Transkription

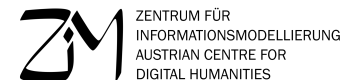
Annotation

Normalisierung

Publikation

<https://digedtnt.github.io>

DigEdTnT Repository



Termine

- 19.09.2023 - FromThePage [Transkription] → ediarum.BASE [Annotation]
- 03.10.2023 - Transkribus [Transkription] → FairCopy [Annotation]
- 17.10.2023 - OpenRefine vs. ba[sic?] [Normalisierung]
- 31.10.2023 - ediarum.BASE [Annotation] → teiPublisher [Publikation]
- 14.11.2023 - FairCopy [Annotation] → ediarum.WEB [Publikation]

jeweils Dienstag, 17:00-18:00 Uhr

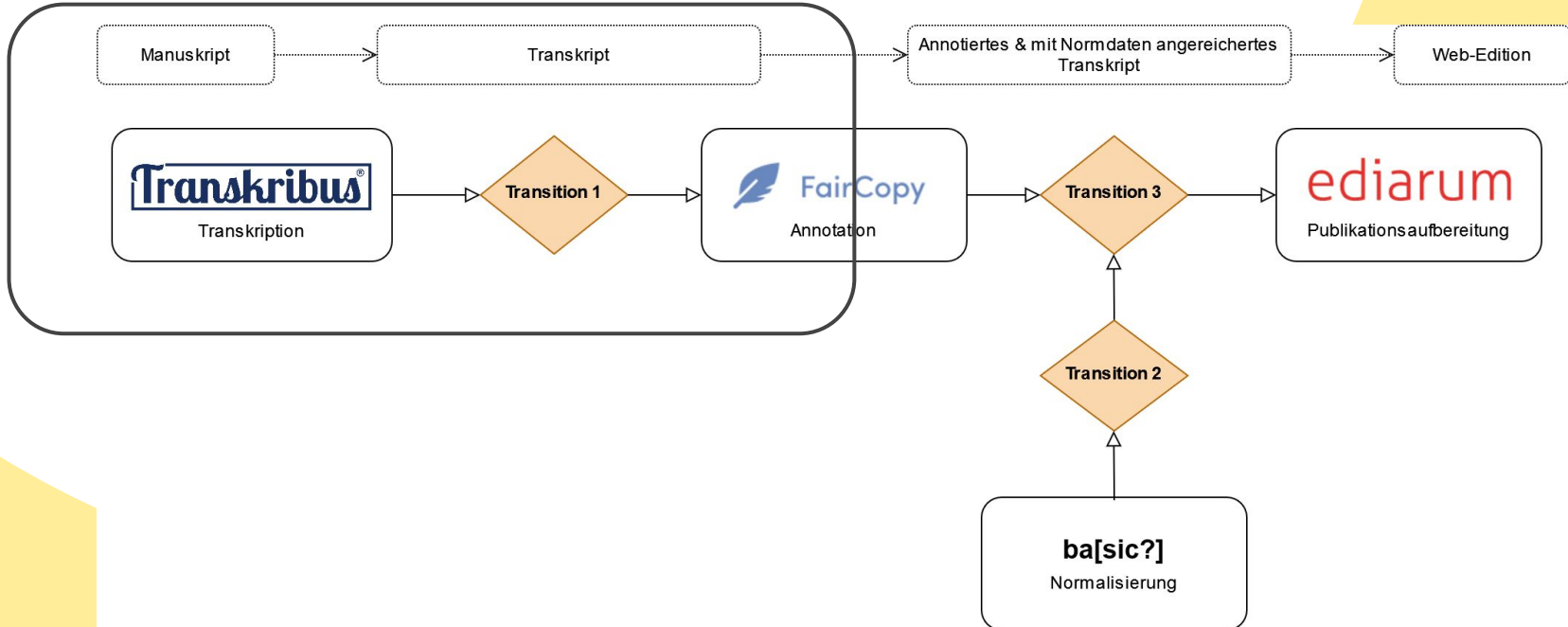
Programm

Tools & Transitions II: Transkribus → FairCopy

- **Pipeline 2:** Beschreibung des Beispielprojekts
- **Transkribus:** Möglichkeiten, Grenzen
- **FairCopy:** Möglichkeiten, Grenzen
- **Transition 2:** Transkribus → XSLT → FairCopy
- **Fragerunde**

Pipeline 2

Datengrundlage: [HSA](#) (Hugo Schuchardt Archiv)
5 verschiedene Briefe, jeweils ~3-7 Seiten



Transkribus



<https://lite.transkribus.eu/>

Tool: Primär Plattform für die automatische KI-gestützte Layout- und Texterkennung von gedruckten und handschriftlichen Texten

Ziel: 1) Automatische Transkription der Briefe Hugo Schuchardts
a) Mittels öffentlicher Modelle
b) Mittels eines mit H. Schuchardts Handschrift trainierten Modells

Zusatz: Grundlegende Annotation von textuellen und strukturellen Merkmale

Kosten: Nur die Texterkennung ist mit Kosten verbunden, alle anderen Funktionalitäten sind kostenfrei

Creditsystem: 1 Credit = 1 handgeschriebene oder 6 gedruckte Seiten
z. B. 300 Credits pro Monat für € 19,90

Transkribus



Möglichkeiten & Anwendungsbereiche

- Keine Installation, keine Hardwarevoraussetzungen - nur Webbrowser nötig
- Transkription und Annotation in Editoransicht mit Bild-Text-Synopse
- Gute Anpassung der Layout- und Texterkennung an die jeweilige Dokumentenstruktur und Hand bzw. Hände über das Training eigener Modelle gegeben
- Selbst definierbare Struktur- und Texttags mit optionalen Attributen anlegbar
- Für kollaborative Nutzung geeignet
- Export in verschiedenen Formaten: TEI-XML, PAGE XML, ALTO XML, PDF (Bild- und Transkriptionslayer), Docx, Tags XLSX, Table XLSX
- Beständige Weiterentwicklung

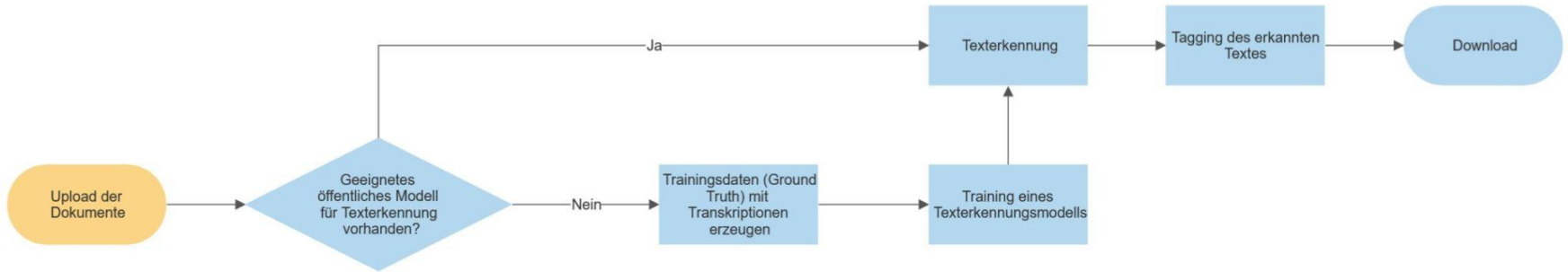
Transkribus



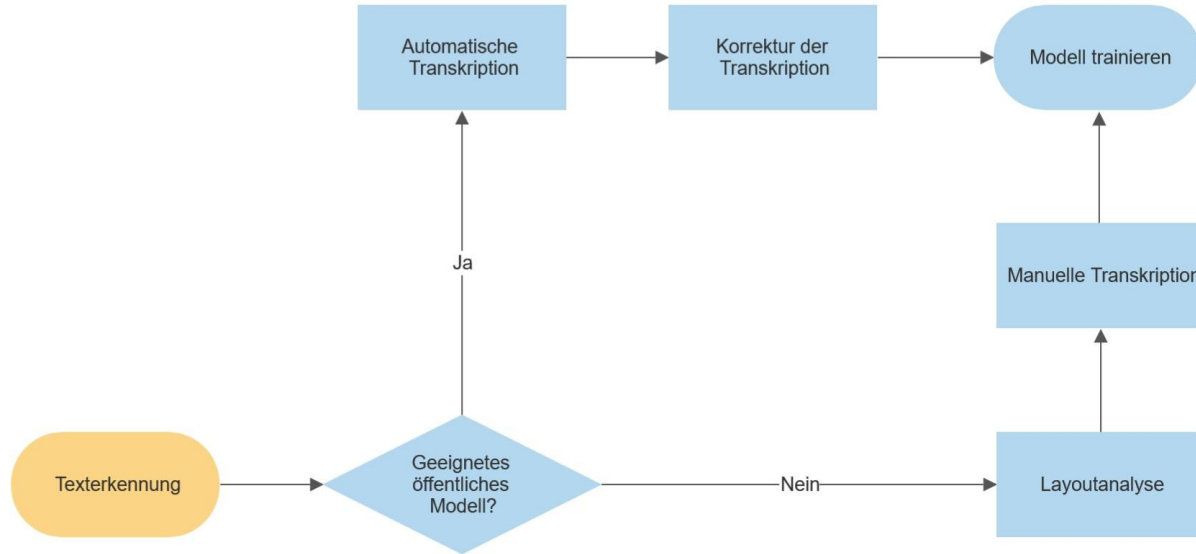
Herausforderungen & Grenzen

- 25-75 transkribierte Seiten (5.000-15.000 Wörter) werden als Trainingsdaten für das Training von Texterkennungsmodellen benötigt
- Editor ist kein vollwertiger Ersatz für ein eigenständiges Annotationstool
- Keine TEI-Validierung im Editor
- TEI-Export nicht valide

Transkribus-Workflow



Transkribus: Transkription



Transkribus: Texterkennung

Text-Erkennung Layout Help

Brief_19_o_hsa.letter.1181

Erkennung starten

Credits needed: -4.00
Available (Personal | Collection) 449 | 0

Smart Search +50% Language Model Advanced Settings

- ★ Favorite Models 0
- 🌐 Public Models 7
- 🔒 Private Models 4

Filter

Search ...

Languages

Deutsch

Deu

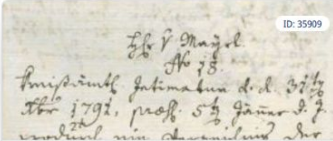
Handschriftlich

Jahrhunderte

1-20

Name	Wörter	Sprache	CER
Search			
The Text Titan I		GER, DUT, FRE, FIN, SWE, ENG	2.95%
The German Giant I	15 420 976	GER	8.30%
Early Kurrent Emperor I.	6 404 094	GER	8.00%
Paul-Goldmann_German-Kurrent_1889-96_V1	63 310	GER	4.40%
Transkribus German handwriting M1	3 610 922	GER	4.70%
German_Kurrent_17th-18th	1 839 841	GER, LA, FR	5.50%
German_Kurrent_XIX_pyala	5 100 439	GER	6.90%

ID: 35909



Public Model

Transkribus German handwriting M1

by Transkribus Community 16/8/2021

Languages GER

Training Set Size 3 610 922

CER (Accuracy) 4.70%

Jahrhunderte 17-20

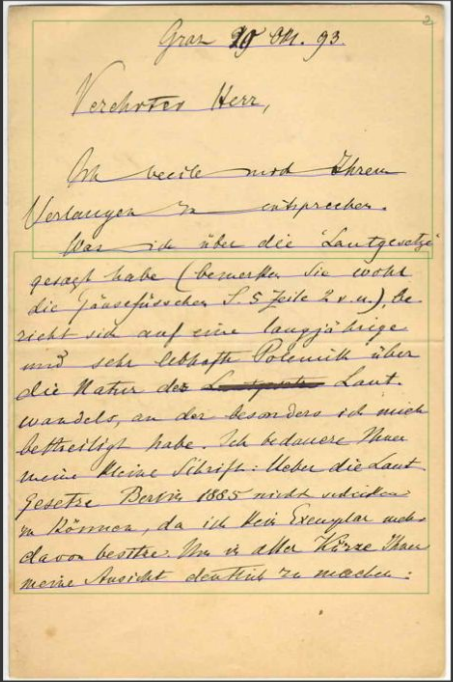
Trained on handwritten

Show Description

Transkribus: Editoransicht

Transkribus Briefe Schuchardt > Brief 1 #1

In Bearbeitung 14.4.2023, 15:53



Das Bild zeigt einen Ausschnitt aus einem handschriftlichen Brief. Der Text ist in cursive geschrieben und beginnt mit 'Graz 29 Okt. 93.' und 'Verehrter Herr,'. Der Haupttext bespricht die Lautgesetze, die in der 'Jäusefüsschen' (S. 5 Zeile 2 v. u.) erwähnt werden. Der Verfasser äußert Bedauern über die Natur des Lantzentens und die kleine Schrift, die in Berlin 1885 nicht schickbar war. Er bittet um die Herstellung einer Kopie.

REGION 1

Graz 29 Okt. 93.
Verehrter Herr.
Ih vecile mod Ohrem
Vverlangen zu entsprechen.
Was ich über die Lautgesetze

REGION 2

gesagt habe (bemerken Sie wohl.
die jäusefüsschen S. 5 Zeite 2 v. u.), be-
riehet sich auf eine laugjährige
und sehr lebhaft Polenik über
die Natur des Lantzenten Laut.
wandels, an der besonders ich mich
betteiligt habe. Ich bedauere Ihnen
meine kleine Schrift: Ueber die Laut.
Gesetze Berlin 1885 nicht schicken
zu können, da ich kein Exemplar mehr
davon besitze. Um in aller Kürze Ihnen
meine Ansicht dentlich zu machen:

Transkribus



Modelltraining

1. Auswahl der Trainingsdaten (mindestens 20 Seiten)
2. Auswahl der Validierungsdaten (10 % automatisch oder manuelle Zuteilung)
3. Angabe der Metadaten (Modellname, Beschreibung, Sprache(n) und Zeitspanne, aus der die Dokumente stammen) und gegebenenfalls

Transkribus



Hugo Schuchardt Handwriting Baseline Transkribus German handwriting M1	10 773	German	2.40%
Schuchardt Handwriting Baseline The German Giant I	10 773	German	2.90%
Hugo Schuchardt Handwriting	10 532	German	5.10%

Transkribus - Annotation

Transkribus®

REGION 1
Graz 19 Okt. 93.
Verehrter Herr,
Ich beeile mich Ihrem
Verlangen zu entsprechen.
Was ich über die ' Lautgesetze'
gesagt habe (bemerken Sie wohl
die Gänsefüsschen S. 5 Zeile 2 v.u.), be-
zieht sich auf eine langjährige
und sehr lebhaft Polemik über
die Natur des Laut-
wandels, an der besonders ich nicht
betheiligt habe. Ich bedauere Ihnen
mein kleine Schrift: Ueber die Laut-
gesetze Berlin 1885 nicht schicken
zu können, da ich kein Exemplar mehr
davon besitze. Um in aller Kürze Ihnen
meine Arbeit deutlich zu machen:

```
<lb facs="#facs_1_tr_4_tl_1" n="N001"/>
<dateline>
  <opener>Graz
    <date when="1893-10-19">19 Okt. 93.</date>
  </opener>
</dateline>
<lb facs="#facs_1_tr_1_tl_1" n="N002"/>
<opener>
  <salute>Verehrter Herr,</salute>
</opener>
```

FairCopy



<https://www.faircopyeditor.com/>

Tool: FairCopy

Ziel: 1) Annotation der textuellen Merkmale
2) Annotation der Named Entities

Kosten: \$ 99 im ersten Jahr, dann \$ 49 für jedes weitere

Möglichkeiten & Anwendungsbereiche

- Transkription und Annotation in Editoransicht
- Kollaborative Transkription und Annotation über die ArchiveEngine-Server von Performant Solutions LLC
- TEI-Validierung - Elemente können nur gemäß TEI-Richtlinien platziert werden
- Qualitätssicherung durch TEI-Validierung und DTA-Basisformat-Schema
- Import von IIIF-Manifesten
- Beständige Weiterentwicklung

Herausforderungen & Grenzen

- Strukturen sind nicht kopierbar, d. h. sich wiederholende Strukturen müssen manuell erneut angelegt werden
- Bei stark verschachtelten Textstrukturpassagen leidet die Übersichtlichkeit
- Bei Markerelementen an Textoberfläche nicht erkennbar, um welche Elemente es sich handelt
- Suchfunktion rudimentär, keine komplexeren Operationen wie Suchen und Ersetzen

Transition: Transkribus → FairCopy

<https://digedtnt.github.io/transition-transkribus-faircopy/>

Tools: Transkribus → FairCopy

Ziel: Valides TEI-Dokument erzeugen, entfernen
nicht benötigter Elemente

→ Transition ≠ generell anwendbares Template
= erweiterbare/erweiterungsbedürftige Grundlage

Transkribus[®]



 FairCopy

Fazit

Transkribus	FairCopy
Benutzerfreundliche Arbeitsumgebung	
Kollaboration	Kollaboration über ArchiveEngine-Server
Annotation einfacher Textstrukturen	Annotation nach TEI-Richtlinien

Ressourcen

- <https://readcoop.eu/de/transkribus/>
- <https://www.faircopyeditor.com/>
- <https://digedtnt.github.io/transkribus/>
- <https://digedtnt.github.io/faircopy/>
- <https://digedtnt.github.io/transition-transkribus-faircopy/>
- https://github.com/DigEdTnT/digedtnt.github.io/tree/master/data/pipelines/pipeline_2

Vielen Dank!

We work for
tomorrow

