



Arbeiten mit Transkribus

Bruno Blüggel
blueggel@uni-greifswald.de

Alle Folien:
„Creative-Commons Namensnennung –
Weitergabe unter gleichen Bedingungen“
CC BY-SA 4.0



1. Kurze Historie
2. Was ermöglicht Transkribus
3. Transkribus Client – Transkribus Web
4. Projektarbeit mit Transkribus
5. Subscriptions-Modelle



Transkribus entstand aus einem EU-Projekt

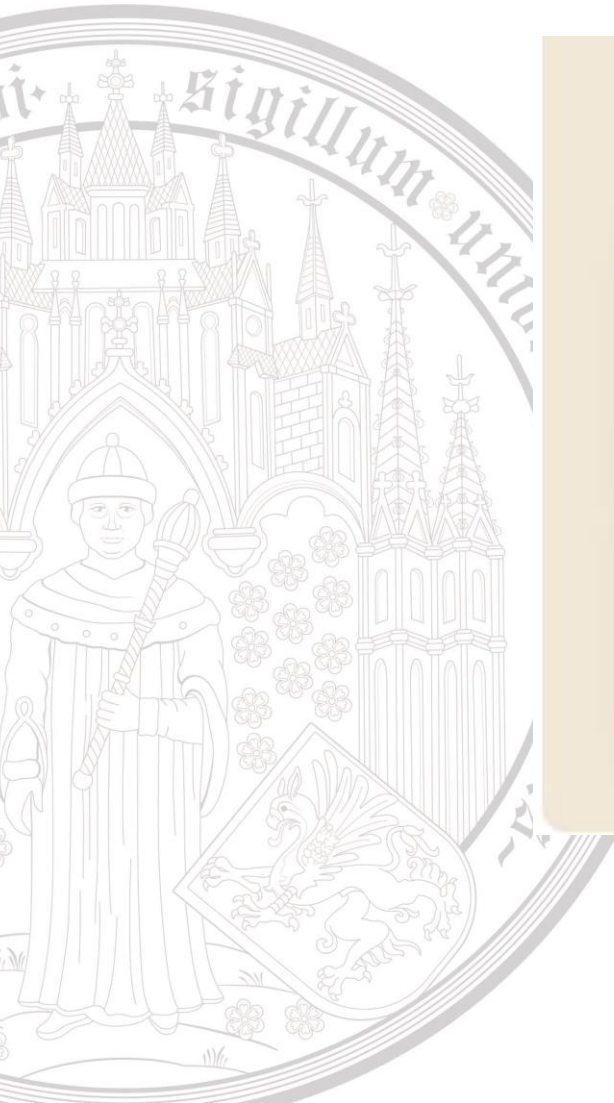
- 2013 tranScriptorium
- 2016 READ
- 2019 Gründung der READ COOP SCE
Uni Greifswald als Gründungsmitglied
- Heute 170 Mitglieder aus 30 Staaten

Wie arbeite ich mit Transkribus?



- Anmelden mit Emailadresse
- Mit 100 freien Credits, kann man starten
- Angemeldete Nutzer können Projekten (Collections) zugeordnet werden
- Collections haben verschiedene Nutzerrechte:
Owner (= Adminrechte) → Leser*in
- Einem Projekt können Credits zugeordnet werden

Transkribus Nutzungs- und Kostenmodelle?



Transkribus Pricing Plans

Individual	Scholar	Organisation
Ideal for Genealogists & Students	Tailored for Individual Researchers	For Research & Cultural Institutions
<ul style="list-style-type: none">AI Text RecognitionCustom AI TrainingDOCX & PDF Export	<ul style="list-style-type: none">Collaboration ToolsAdvanced AI ToolsTranskribus Sites	<ul style="list-style-type: none">User ManagementDedicated Success ManagerAPI Access

41,90 €

77,90 €

Individueller Mitgliedsbeitrag

Was ermöglicht Transkribus?



- Manuelle und automatisierte Transkription von handschriftlichen und gedruckten Dokumenten (und „Mischformen“ !)
- Training von KI - Modellen
- Zusammenarbeit in Editionsteams (Collection)
- Nutzung einer mächtigen Volltextsuche in Dokumenten
- Tagging von Strukturelementen und Inhalten
- Export der Ergebnisse in verschiedenen Datenformaten

Was braucht Transkribus?



- Welche Regionen einer Seite sollen bearbeitet werden? → Textregion → Layoutanalyse
- Abfolge der einzelnen Textregionen
- Zeilen: Anfang – Ende – Oberlängen – Unterlängen
- Trainingsmaterial → „Ground Truth“

Ausgangsmodell → ausreichend oder weiter entwickeln?



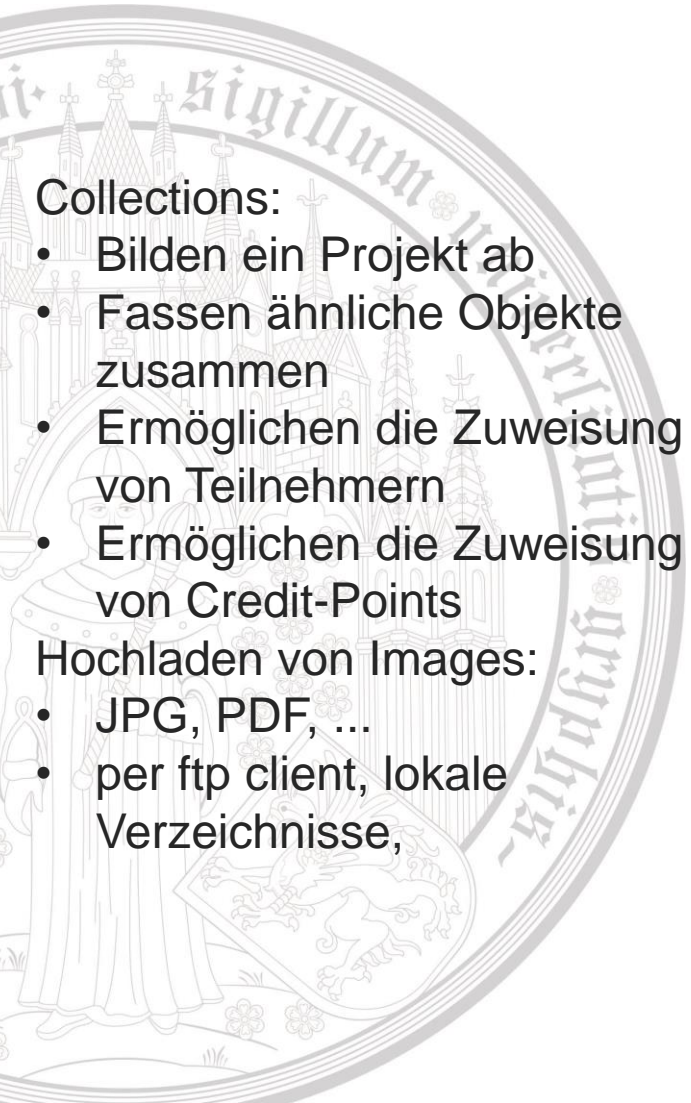
Transkribus Client – ist das „historische“ Basisprogramm

Vorteile:

- Vergleichsweise schnell und stabil
- Gute Projektsteuerungs- und Verwaltungstools

Nachteile:

- Benötigt Java → Installation und update sind aufwendiger → höhere Einstiegshürde
- Die eigentliche Transkribierungs-Oberfläche ist etwas komplizierter und wirkt „angestaubt“



Collections:

- Bilden ein Projekt ab
- Fassen ähnliche Objekte zusammen
- Ermöglichen die Zuweisung von Teilnehmern
- Ermöglichen die Zuweisung von Credit-Points

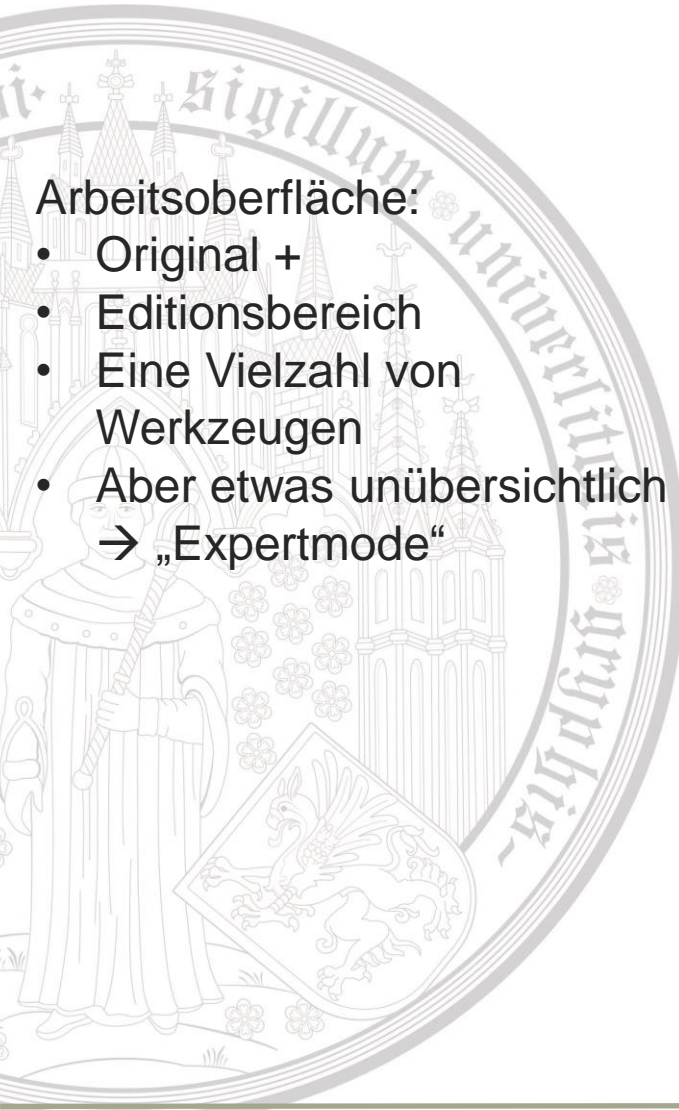
Hochladen von Images:

- JPG, PDF, ...
- per ftp client, lokale Verzeichnisse,

The screenshot shows the Transkribus v1.17.0 client interface. The main window displays a document viewer with musical notation and a 'Document ingest / upload' dialog box. A table of collections is also visible in the foreground.

ID	Name	Role	Description
39852	Wismarer Tribunal	Editor	created by alverman@uni-greifswald.de
41600	Transkribus workshop	Editor	created by guenter
43485	Volksliedarchiv-Noten	Transcriber	created by alverman@uni-greifswald.de
44082	Niederdeutsch	Owner	created by alverman@uni-greifswald.de
78485	Pomerania Monastica Production	Owner	created by alverman@uni-greifswald.de
123676	Garnison	Editor	created by psettgast@wismar.de
125699	Mikrofilme	Editor	created by alverman@uni-greifswald.de
146376	EOOPEN	Owner	created by blueggel@uni-greifswald.de
151864	Volkslied_Mappen_Layout	Transcriber	created by alverman@uni-greifswald.de
152313	Theologie	Owner	created by blueggel@uni-greifswald.de
206317	Queen Charlottes Letters	Owner	created by blueggel@uni-greifswald.de
210087	Neustrelitz	Owner	created by blueggel@uni-greifswald.de
212463	Dalman	Owner	created by blueggel@uni-greifswald.de
215543	Volkslied_Liederbücher_Production	Owner	created by josts@uni-greifswald.de
215544	Volkslied_PLM_1_Production	Editor	created by josts@uni-greifswald.de
215545	Volkslied_PLM_2_Production	Owner	created by josts@uni-greifswald.de
217063	Wismar_Kirchenbuecher	Owner	created by blueggel@uni-greifswald.de
2231...	Opera	Owner	created by blueggel@uni-greifswald.de
231083	Ratsprotokolle_HST	Owner	created by blueggel@uni-greifswald.de

The 'Document ingest / upload' dialog box shows options for uploading documents: 'Upload via private FTP (also PDF files)', 'Upload via URL of DFG Viewer METS', and 'Extract and upload images from pdf'. The 'Upload single document' option is selected. The 'Local folder' and 'Title on server' fields are empty. The 'Add to collection' dropdown is set to 'Opera (223157, Owner)'. The 'Upload' and 'Cancel' buttons are visible at the bottom.

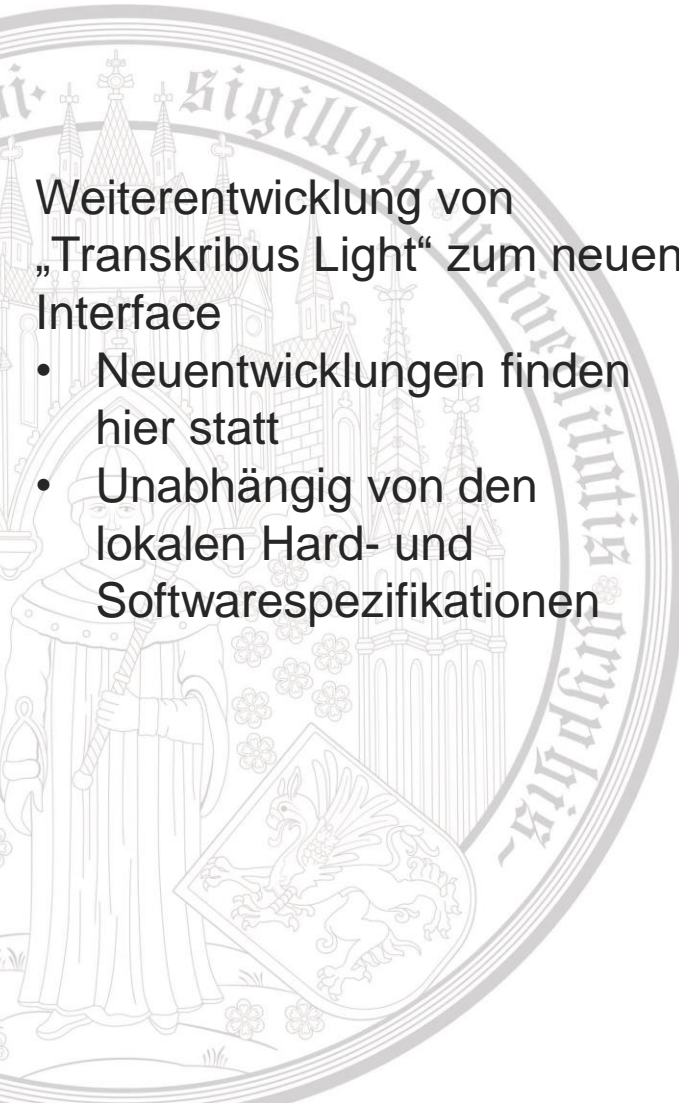


Arbeitsoberfläche:

- Original +
- Editionsbereich
- Eine Vielzahl von Werkzeugen
- Aber etwas unübersichtlich
→ „Expertmode“

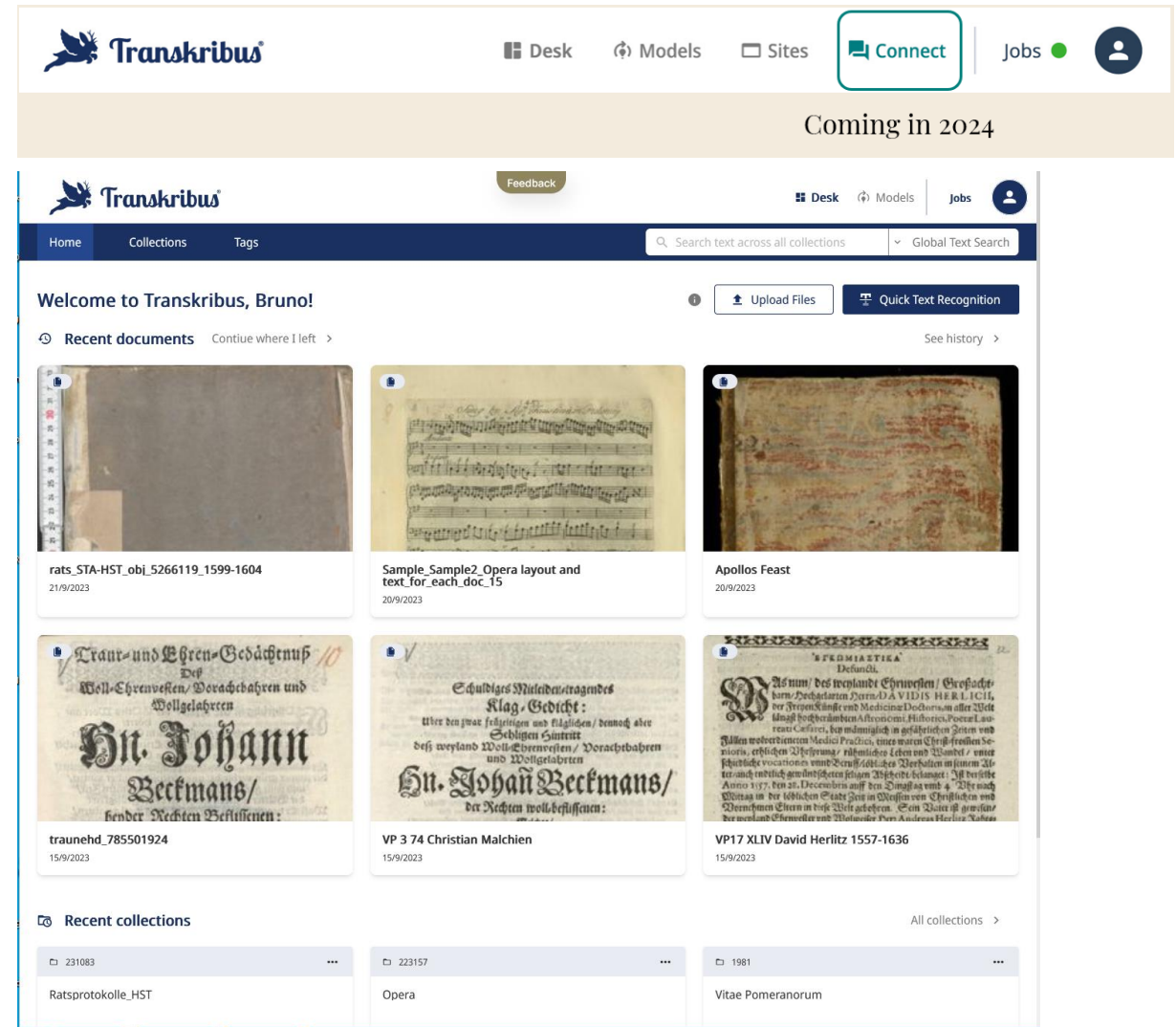
The screenshot shows the Transkribus-Client interface. On the left, there is a 'Documents' table with columns for ID, Title, Pages, Uploader, Uploaded, and Collections. The table lists various documents, with the selected document being 'Sample_Sample2_Opera layout and text_for_each_doc_15'. On the right, the main view displays a musical score with lyrics. The score is for 'Sung by Sig.^{ra} Faustina in Ptolomey' and includes the tempo marking 'Andante'. The lyrics are: 'Quanto è felice quel'augel-letto che senza pene servà volando così goden--do sua liber--'. Below the score, there is a list of annotations for the lyrics, such as '1-1 2', '2-1 45', '3-1 Sung by Sig.^{ra} Faustina in Ptolomey', etc.

ID	Title	Pages	Uploader	Uploaded	Collections
15793...	Sample_Fields test_for_each_doc_2	4	r.ribuoli@rea...	Tue Sep 19 10:...	(Opera,223157)
1568...	Sample_Sample2_Opera layout and t...	90	r.ribuoli@rea...	Tue Sep 12 1...	(Opera,22315...
15594...	Sample_Baseline Recognition_sample ...	96	r.ribuoli@rea...	Tue Sep 05 10:...	(Opera,223157)
15367...	ouve_3821_1_Noten_Schwerin	50	blueggel@uni...	Fri Aug 18 15:4...	(Opera,223157)
15366...	concyr_2435_1_Noten_Schwerin	154	blueggel@uni...	Fri Aug 18 12:1...	(Opera,223157)
15366...	apofte_182_1_Noten_Schwerin	196	blueggel@uni...	Fri Aug 18 11:4...	(Opera,223157)
15366...	thfr_25262_Noten_Schwerin	360	blueggel@uni...	Fri Aug 18 11:2...	(Opera,223157)
15366...	sifru_2439_1_Noten_Schwerin	68	blueggel@uni...	Fri Aug 18 11:1...	(Opera,223157)
15366...	otho_2409_1_Noten_Schwerin	98	blueggel@uni...	Fri Aug 18 11:1...	(Opera,223157)
15359...	rosecco_3818_1_Noten_Schwerin	171	blueggel@uni...	Fri Aug 17 17:...	(Opera,223157)
15359...	catfmu_1329_1_3_3_Noten_Schwerin	324	blueggel@uni...	Thu Aug 17 16:...	(Opera,223157)
15359...	apofteort_2407_3_1_1_Noten_Schwerin	254	blueggel@uni...	Thu Aug 17 16:...	(Opera,223157)
15358...	airi_3822_1_Noten_Schwerin	24	blueggel@uni...	Thu Aug 17 16:...	(Opera,223157)
15346...	themoces_2401_1_Noten_Schwerin	32	blueggel@uni...	Wed Aug 16 1:...	(Opera,223157)
15346...	libepraf_2407_2_Noten_Schwerin	60	blueggel@uni...	Wed Aug 16 1:...	(Opera,223157)
15346...	catfmu_1329_1_1_1_Noten_Schwerin	314	blueggel@uni...	Wed Aug 16 1:...	(Opera,223157)
15317...	Sample_Layout Recognition_sample 1...	96	r.ribuoli@rea...	Mon Aug 14 1:...	(Opera,223157)
15265...	apofteort_2407_3_4	284	blueggel@uni...	Tue Aug 08 17:...	(Dalman,2124...
15265...	apofteort_2407_3_3	224	blueggel@uni...	Tue Aug 08 17:...	(Opera,223157)
15264...	apofteort_2407_3_2	250	blueggel@uni...	Tue Aug 08 16:...	(Opera,223157)
15264...	Apollos Feast	254	blueggel@uni...	Tue Aug 08 16:...	(Opera,223157)



Weiterentwicklung von „Transkribus Light“ zum neuen Interface

- Neuentwicklungen finden hier statt
- Unabhängig von den lokalen Hard- und Softwarespezifikationen



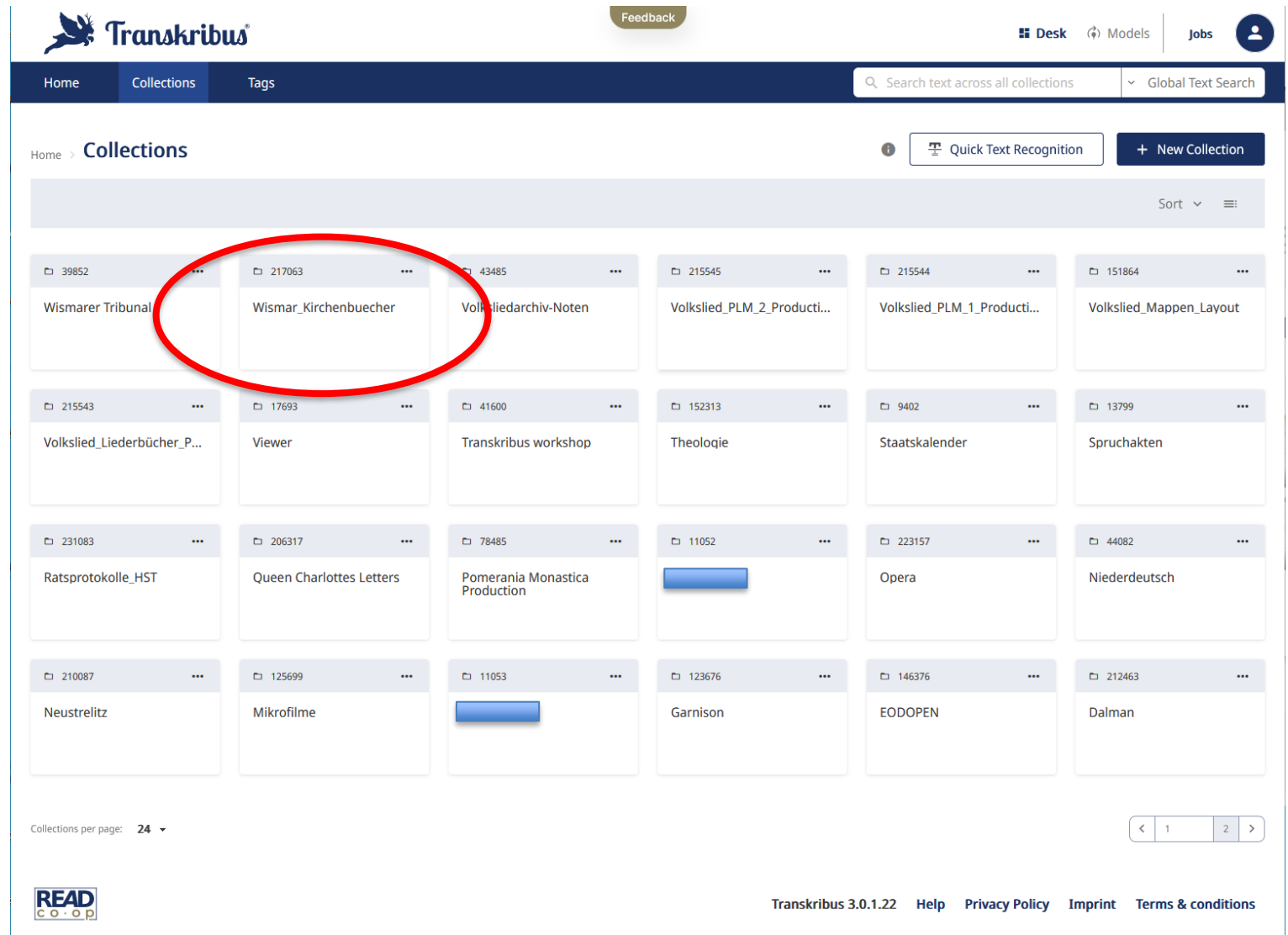
The screenshot shows the Transkribus web interface. At the top, there is a navigation bar with icons for Desk, Models, Sites, Connect, Jobs, and a user profile. Below this is a banner that says "Coming in 2024". The main content area has a "Welcome to Transkribus, Bruno!" message and a "Recent documents" section. This section displays a grid of document thumbnails with their titles and upload dates. The documents shown are:

- rats_STA-HST_obj_5266119_1599-1604 (21/9/2023)
- Sample_Sample2_Opera layout and text_for_each_doc_15 (20/9/2023)
- Apollos Feast (20/9/2023)
- traunehd_785501924 (15/9/2023)
- VP 3 74 Christian Malchien (15/9/2023)
- VP17 XLIV David Herlitz 1557-1636 (15/9/2023)

Below the documents, there is a "Recent collections" section showing a grid of collection thumbnails with their names and IDs:

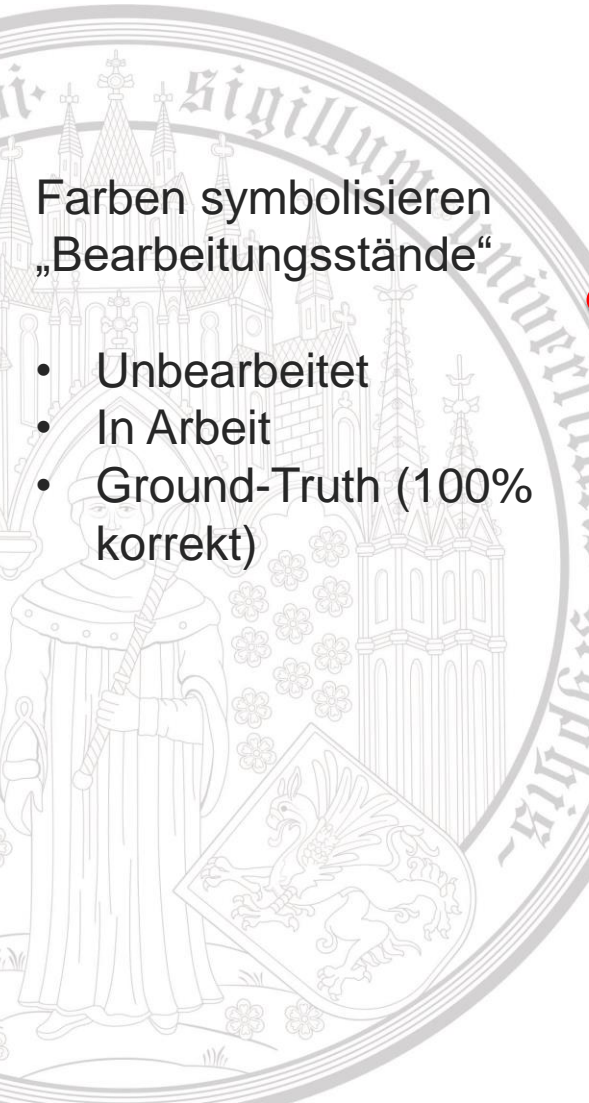
- 231083 Ratsprotokolle_HST
- 223157 Opera
- 1981 Vitae Pomeranorum

Transkribus-Web: Collections



The screenshot shows the Transkribus web interface. At the top, there is a navigation bar with 'Home', 'Collections', and 'Tags' tabs. A search bar is present with the text 'Search text across all collections'. Below the navigation bar, the main content area displays a grid of collection cards. Each card shows a folder icon, a collection ID, and the collection name. The collection 'Wismar_Kirchenbuecher' with ID 217063 is circled in red. Other visible collections include 'Wismarer Tribunal', 'Volksliedarchiv-Noten', 'Volkslied_PLM_2_Producti...', 'Volkslied_PLM_1_Producti...', 'Volkslied_Mappen_Layout', 'Volkslied_Liederbücher_P...', 'Viewer', 'Transkribus workshop', 'Theologie', 'Staatskalender', 'Spruchakten', 'Ratsprotokolle_HST', 'Queen Charlottes Letters', 'Pomerania Monastica Production', 'Opera', 'Niederdeutsch', 'Neustrelitz', 'Mikrofilme', 'Garnison', 'EODOPEN', and 'Dalman'. At the bottom of the page, there is a footer with the 'READ' logo, the version 'Transkribus 3.0.1.22', and links for 'Help', 'Privacy Policy', 'Imprint', and 'Terms & conditions'.

Transkribus-Web: Collections



Farben symbolisieren
„Bearbeitungsstände“

- Unbearbeitet
- In Arbeit
- Ground-Truth (100% korrekt)



Feedback

Desk Models Jobs

Home Collections Tags

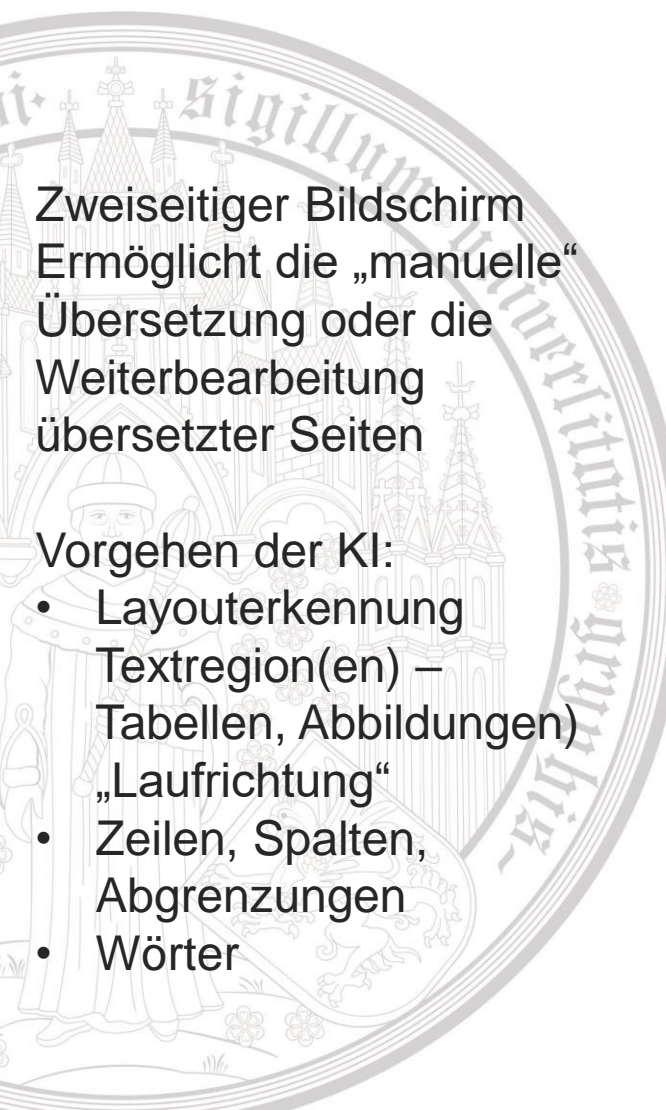
Search text across all collections Global Text Search

Home > Collections > Wismar_Kirchenbuecher > 04hege_PPN_ST...obj_5631403_37

0 Selected Recognize Train Model

Pages Filter (0)

Pages per page: 48



Zweiseitiger Bildschirm
Ermöglicht die „manuelle“
Übersetzung oder die
Weiterbearbeitung
übersetzter Seiten

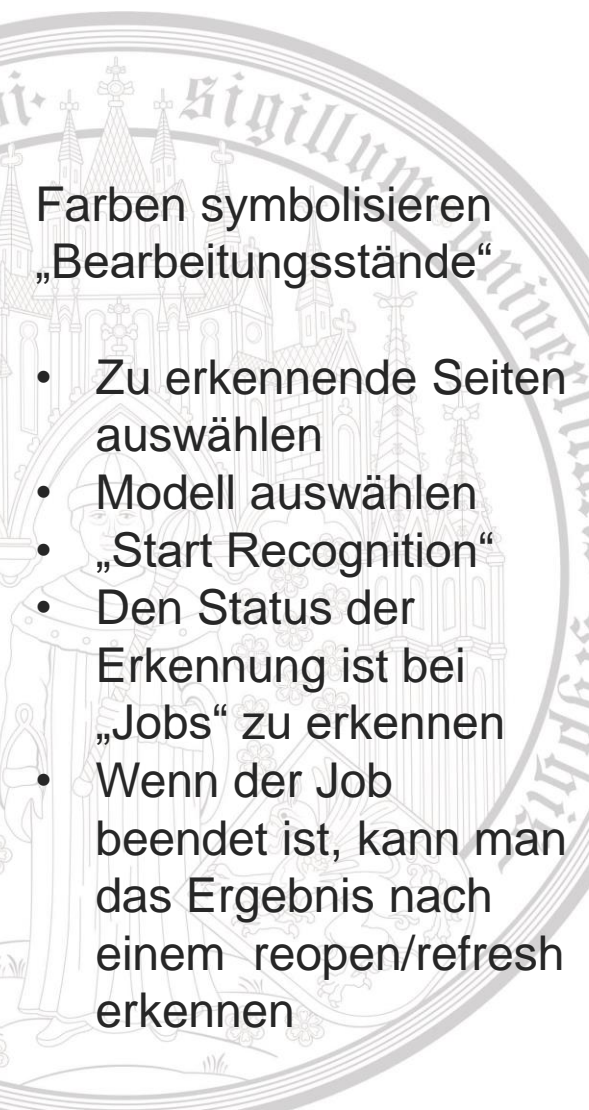
Vorgehen der KI:

- Layouterkennung
Textregion(en) –
Tabellen, Abbildungen)
- „Laufrichtung“
- Zeilen, Spalten,
Abgrenzungen
- Wörter



- Anmelden mit Emailadresse
- Collection anlegen 
- Dokumente hochladen 
Images (jpg oder png; max. 10 MB / Image)
PDF
- Modell auswählen (derzeit 150 public models)
Es gibt inzwischen zwar sehr mächtige Modelle (Supermodels),
aber trotzdem: je diffiziler die Vorlagen, desto differenziertere
Modelle sollte man für seine spezifischen Aufgaben trainieren
- Wenn man sein Modell frei gibt, entsteht ein weiteres public model

Text-Erkennung



Farben symbolisieren „Bearbeitungsstände“

- Zu erkennende Seiten auswählen
- Modell auswählen
- „Start Recognition“
- Den Status der Erkennung ist bei „Jobs“ zu erkennen
- Wenn der Job beendet ist, kann man das Ergebnis nach einem reopen/refresh erkennen

Transkribus

Feedback

Desk Models **Jobs**

Home Collections Tags

Search text across all collections Global Text Search

Home > Collections > Wismar_Kirchenbuecher > 04hege_PPN_ST...obj_5631403_37

0 Selected Recognize Train Model

Pages Filter (0)

Pages per page: 48

Transkribus – 150 Public models



Modell auswählen

- Handschrift
- Gedruckt
- Handschrift / Gedruckt
- Größe des Trainingssets und die Character Error Rate CER beachten

The screenshot shows the Transkribus interface with the following elements:

- Search bar: "D - Peabody Newspaper Index Cards"
- Buttons: "Text Recognition", "Layout", "Smart Search +50%", "Language Model", "Advanced Settings", "Start Recognition"
- Table of models:

NAME	TRAINING SET SIZE	LANGUAGE
The German Giant I	15 420 976	GER
The Dutchess I	11 693 499	DUT
Transkribus Print M1	5 068 310	GER, ENG, DUT,
Transkribus French Model 1	1 933 011	FRE
Transkribus English Handwriting M3	2 125 253	ENG

Filter: Handwritten or Printed

Model details for "The German Giant I":

- by Transkribus Community
- 20/3/2023
- Languages: GER
- Training Set Size: 15 420 976
- CER (Accuracy): 8.30%



Bevor ich ein Modell trainiere, sollte ich folgende Dinge bedenken:

- Wie umfangreich sind meine Texte (ein Brief, Postkartensammlung, Tagebuch, komplette handschriftliche Aktensammlungen)
- Welchen Zeitbereich umfassen meine Texte
Akten über mehrere Jahrhunderte?
Briefe einer Verfasser*in (der Schreibstil ändert sich im Laufe des Lebens)
- Entsprechend muss ich mein Trainingsmaterial aussuchen
- Welches Ziel habe ich?
Volltextsuche – absolut fehlerfreie Edition
- Wie gehe ich mit Abkürzungen um?
Auflösen – belassen – taggen
- Genügt ein mächtiges Standardmodell + Korrektur?
Oder trainiere ich es weiter?



Die Performance eines Modells hängt ab:

- Von der Quantität der zur Verfügung gestellten Trainings- und Validierungsdaten
- Vom Unterschied zwischen einem vollständig und korrekt übersetzten Text (**Character Error Rate CER** = Prozentsatz der Zeichen, die falsch übersetzt wurden)
- Die Performance kann im Einsatz verbessert werden durch verbesserte Sucheinstiege (Smart Search oder unscharfe Suche)
- Oder den Einsatz eines spezifischen Sprachmodells (z.B. „Niederdeutsch 16. Jh.“)

Tagging

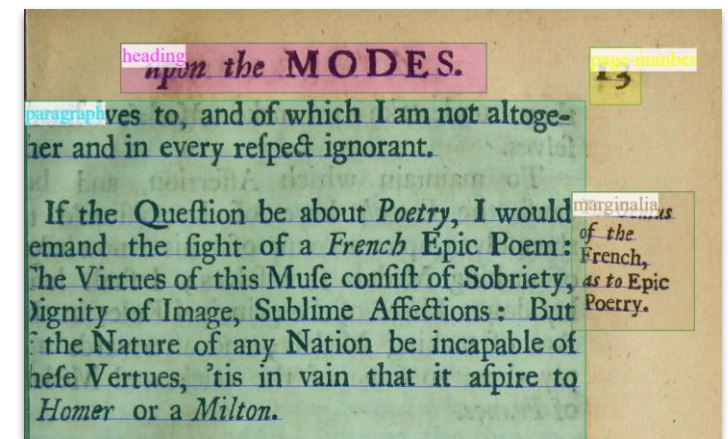
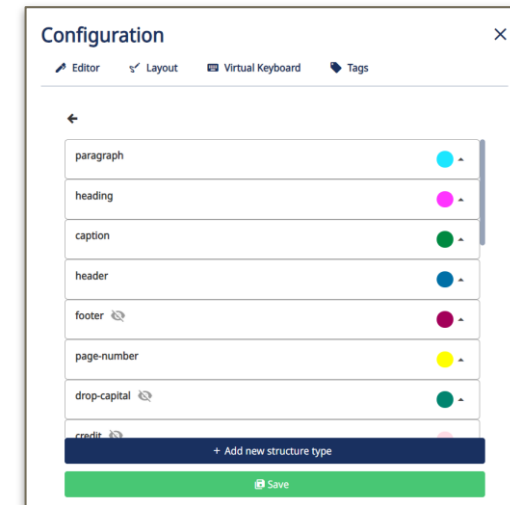


Structural Tags: zur Markierung struktureller Merkmale
Im Editor sind eine Anzahl vorgefügter Tags vorhanden
neue können editiert werden

ACHTUNG:

Wenn Sie neue Tags generieren, denken Sie immer an die
Weiterverwendung in anderen Programmen.

Ein perfektes individuelles Modell mag zwar für eine spezielle
Aufgabe (Editionsprojekt) optimal sein – aber immer an den
Datenaustausch denken (z.B. Verlag)

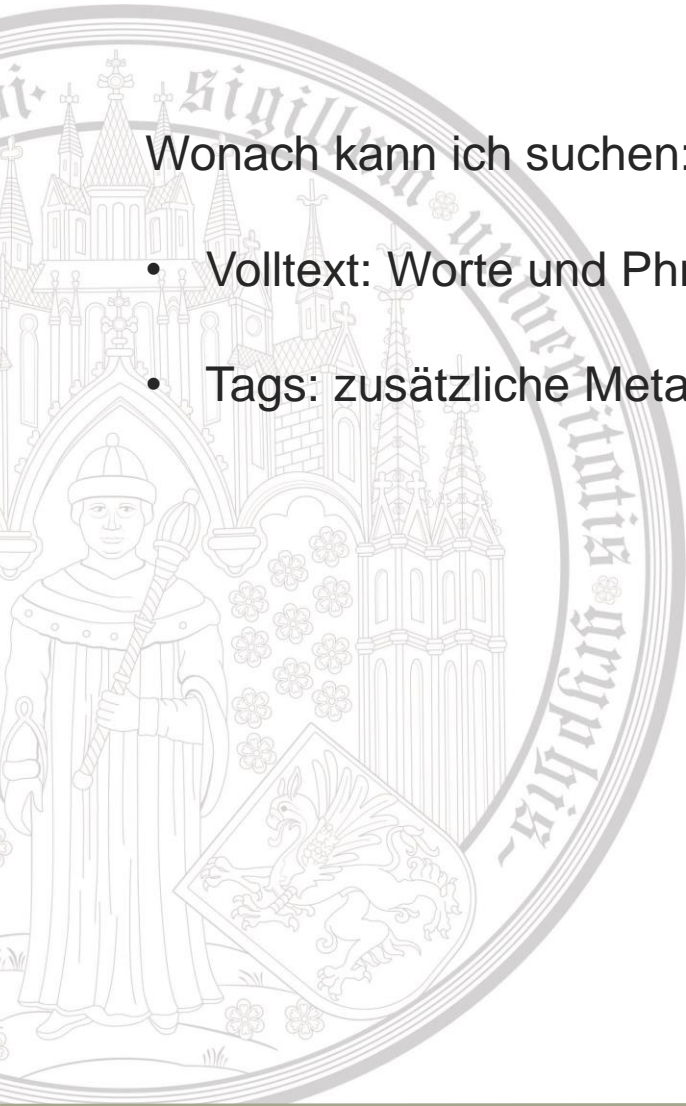


Suche



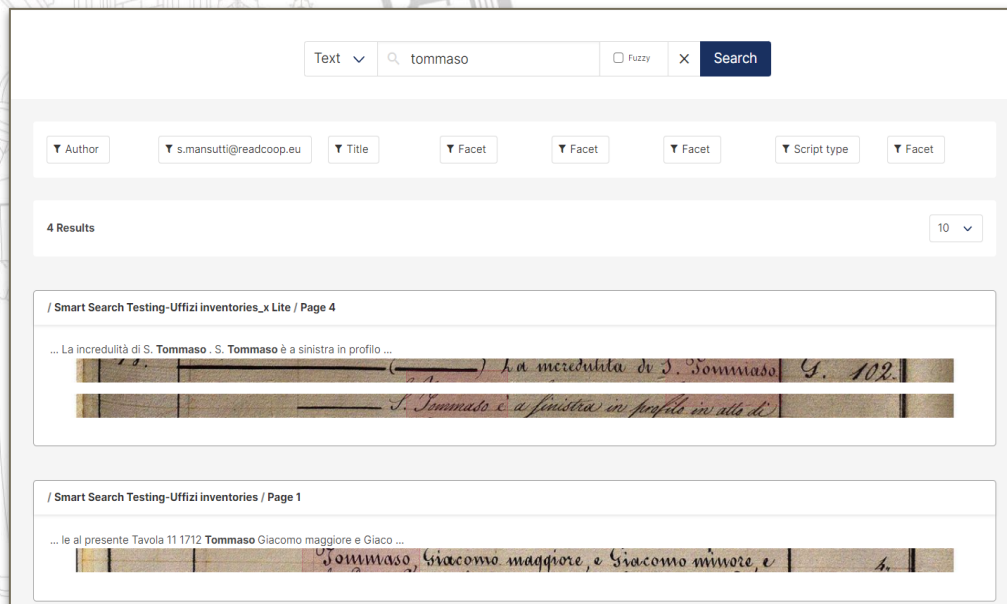
Wonach kann ich suchen:

- Volltext: Worte und Phrasen + unscharfe Suche (Smart Search)
- Tags: zusätzliche Metadaten, die als Tags eingegeben wurden

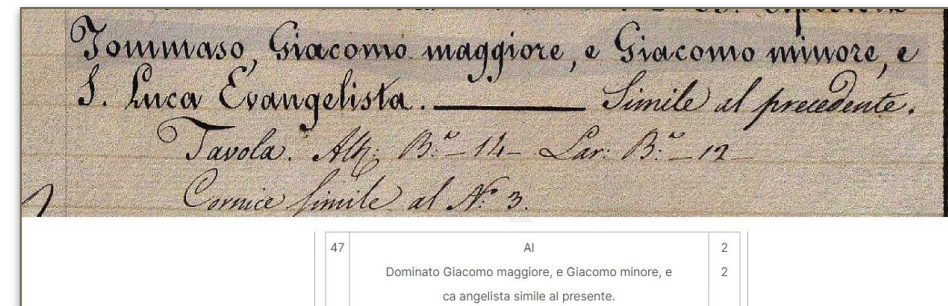


Smart Search

- Smart Search = unscharfe Suche
- Beispiel Tommaso wird bei einer CER von 20% auch gefunden, wenn es falsch interpretiert wurde („Dominato“)
- Die Funktion „Smart Search“ muss beim Übersetzen durch ein Flag gesetzt werden



The screenshot shows a search interface with a search bar containing 'tommaso' and a 'Search' button. Below the search bar are several facets: Author (s.mansutti@readcoop.eu), Title, and three Facet buttons. The results section shows 4 results. The first result is titled 'Smart Search Testing-Uffizi inventories_x Lite / Page 4' and contains a snippet of text: '... La incredulità di S. Tommaso. S. Tommaso è a sinistra in profilo ...'. Below this snippet are two image thumbnails of manuscript pages. The second result is titled 'Smart Search Testing-Uffizi inventories / Page 1' and contains a snippet of text: '... le al presente Tavola 11 1712 Tommaso Giacomo maggiore e Giaco ...'. Below this snippet is one image thumbnail of a manuscript page.



The image shows a handwritten manuscript snippet in cursive script. The text reads: 'Tommaso, Giacomo maggiore, e Giacomo minore, e S. Luca Evangelista. Simile al precedente. Tavola. Alf. B. 11- Lar. B. 12. Cornice simile al N. 3.' Below the snippet is a digital transcription table.

47	Al	2
	Dominato Giacomo maggiore, e Giacomo minore, e ca angelista simile al presente.	2

Exportfunktion



Images

- METS (Metadata Encoding and Transmission Standard) file
- Images

Page XML

- Zusatzinformationen über digitalisierte Seite (Layoutstrukturen, Seiteninhalt)

ALTO

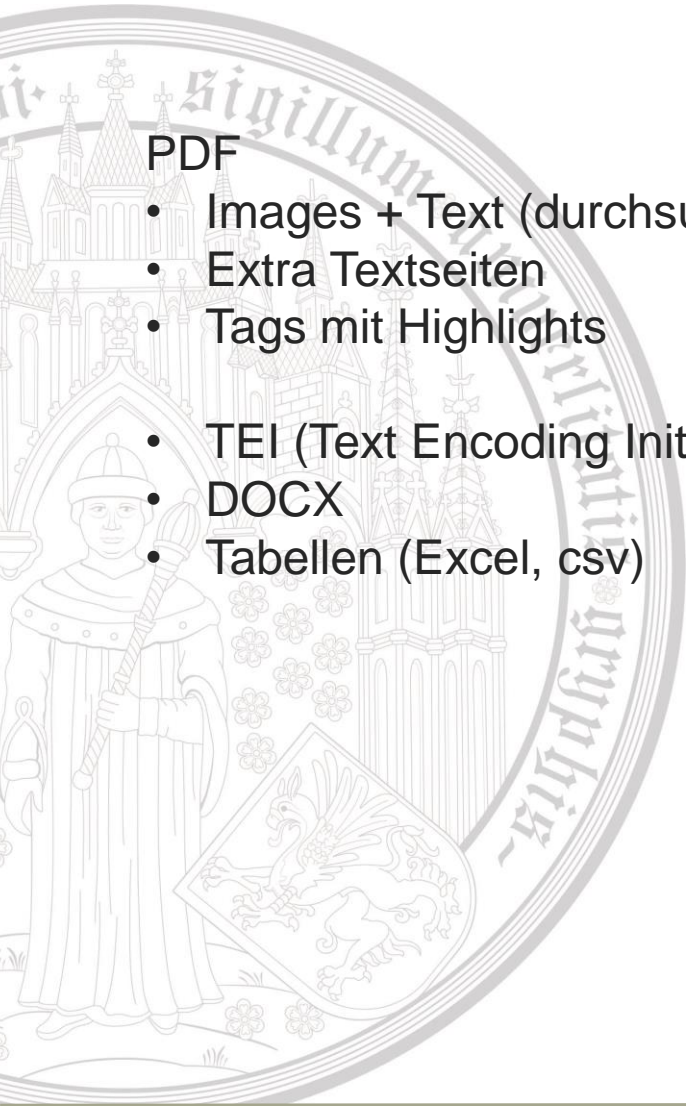
- XML Schema mit technischen Metadaten (Layout, Wordpositionen)

Exportformate



PDF

- Images + Text (durchsuchbares PDF)
- Extra Textseiten
- Tags mit Highlights
- TEI (Text Encoding Initiative)
- DOCX
- Tabellen (Excel, csv)



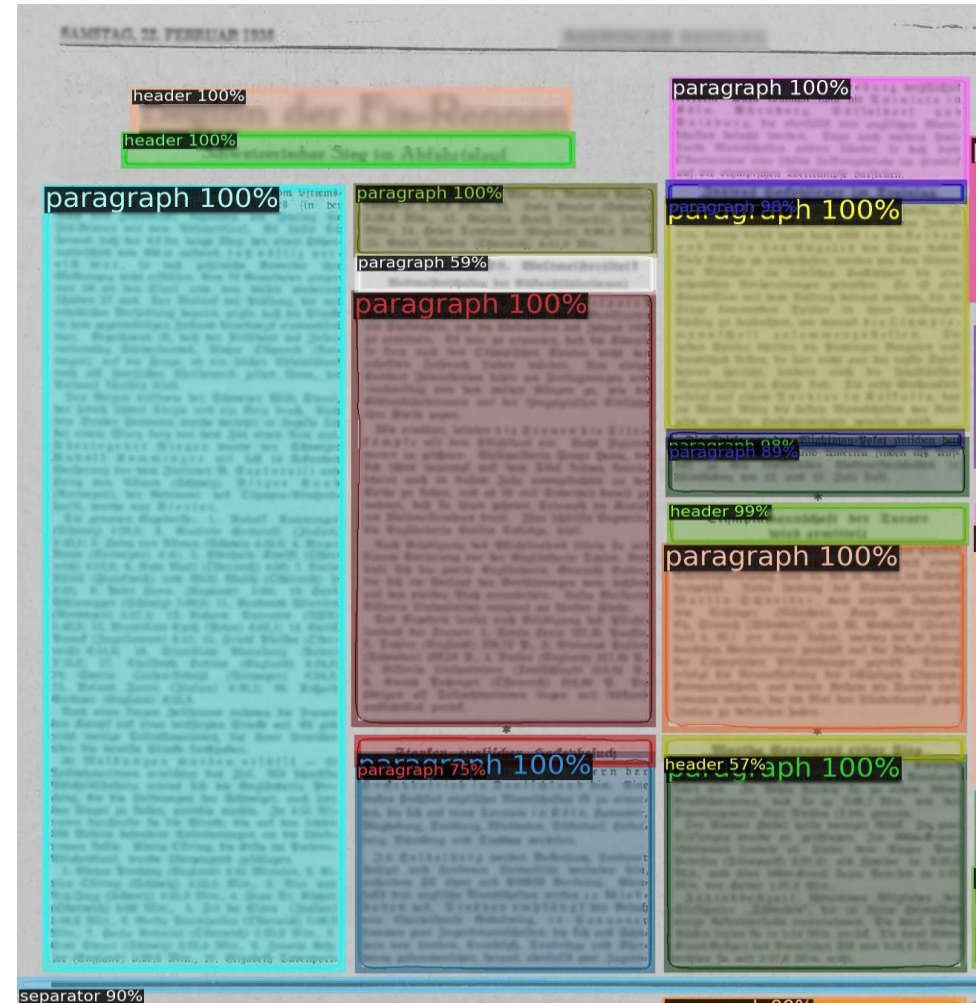
Baustellen = Beta Versionen 😊

Field Models (Beta Version)

Bisher ist beim Trainieren der
Layouterkennung noch viel „händische“

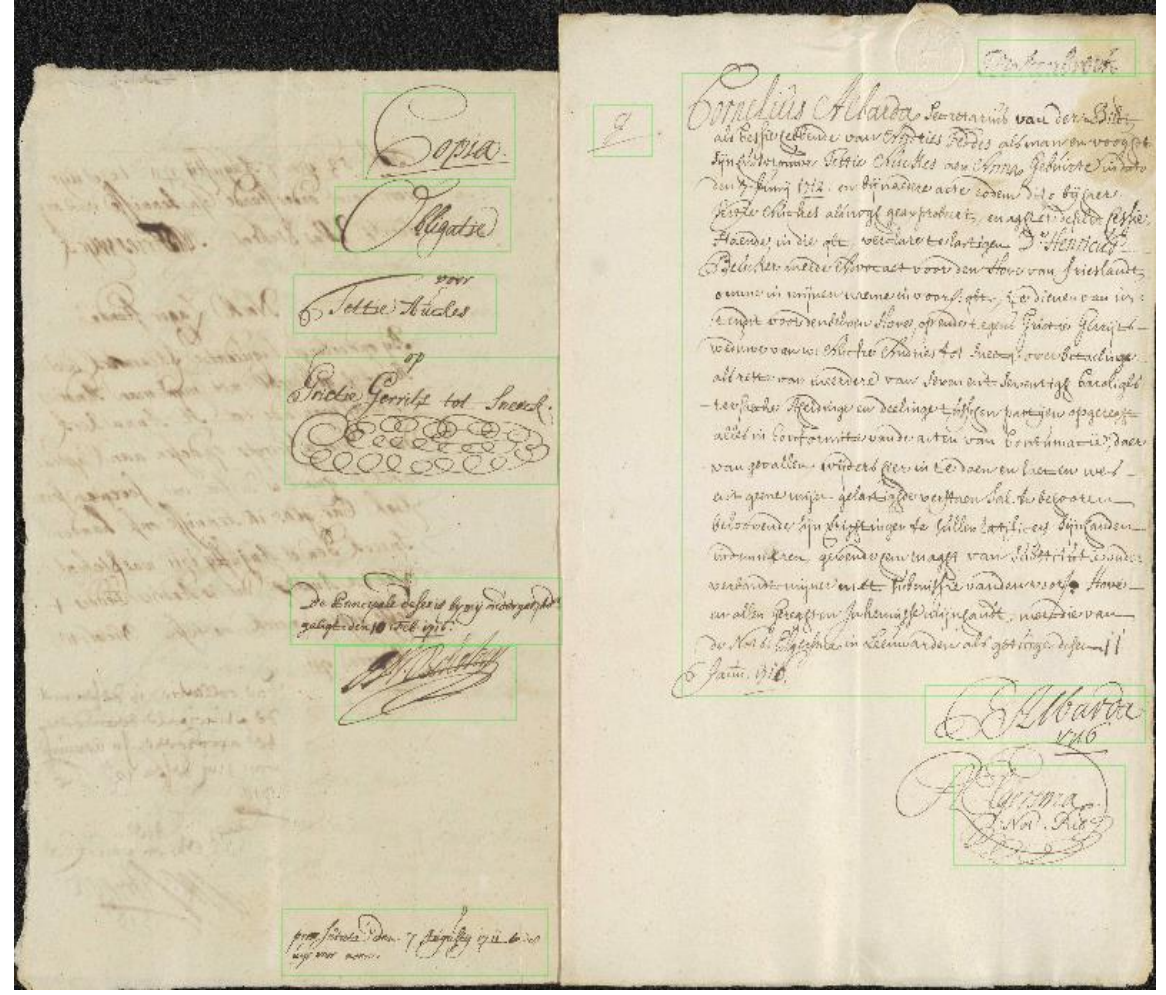
Arbeit erforderlich

Automatisierte Layouterkennung für
Zeitungssegmentierung



Baustellen = Betaversionen 😊

Automatisierte Layouterkennung für Text Regionen



Baustellen = Betaversion 😊

Automatisierte Layouterkennung für Formaterkennung



alk

Vater: } **vater_separiert 100%** *1957*
Mutter: } **mutter_separiert 100%** *30*

Staatsangehörigkeit: **Staatsangehörigkeit 99%**

Personalakt: Familienname: **name 100%**
Vorname: **vorname 99%**

Geburtsdatum: **datum 100%** Geburtsort: **ort 99%**

Glaubensbek.: **Religion 100%** Kreis: *prov.*

Beruf: 1. **Beruf 100%** 3.
4. 5. 6.

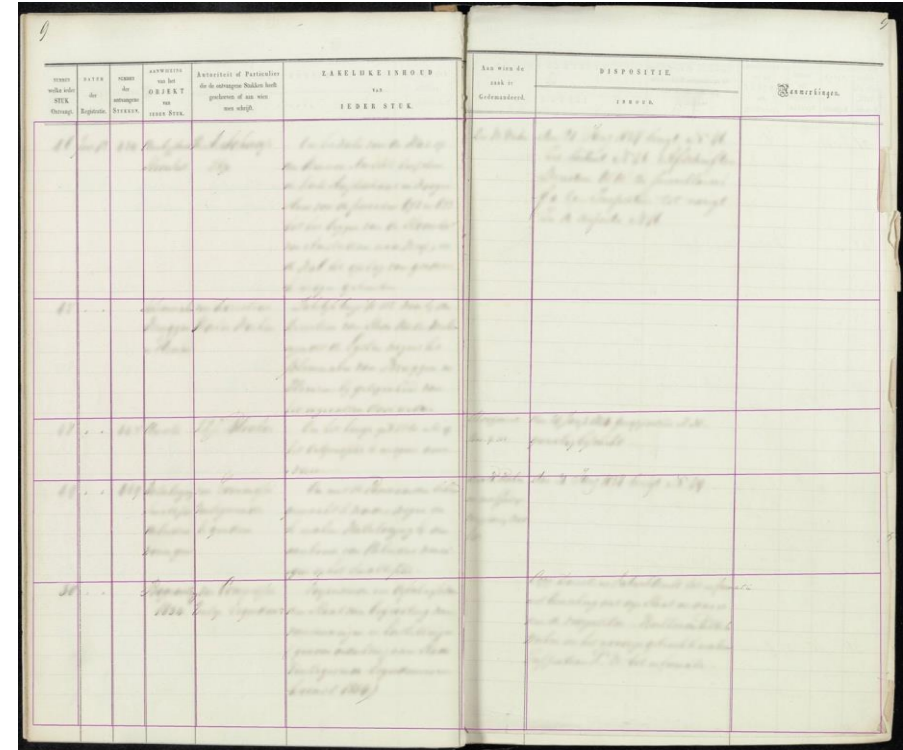
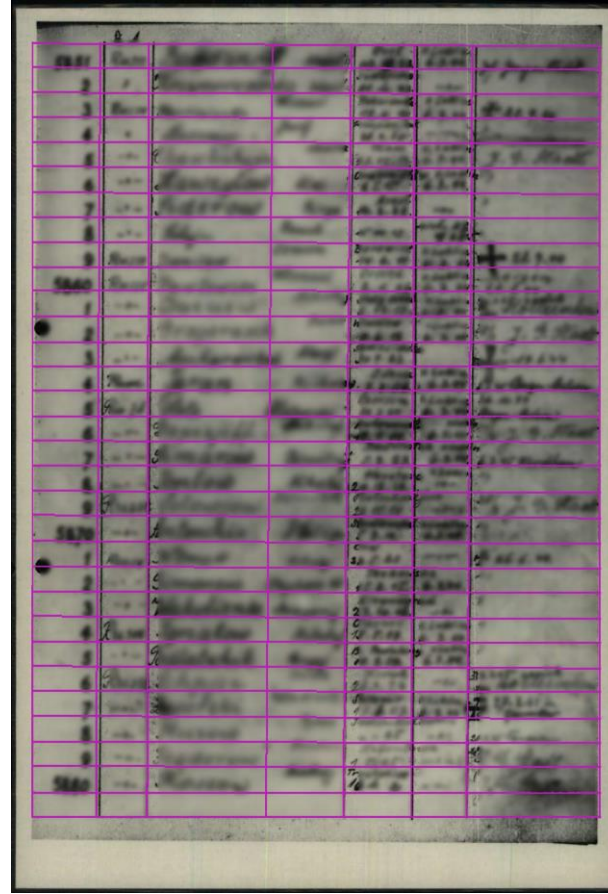
Familienangehörige	Geburts- tag mo- nat	jahr	Geburtsort (Kreis, Provinz) Standesamt	Glaubens- bek.	Aus- zugs- verm.	Mitglied und Stellung i. d. NSDAP, oder einer ihrer Gliederungen	Vermerke
Kinder:							
Verheiratet seit verheiratet(u0020seit 99% Standesamt Standesamt 97%							
Standesamtsnummer 99% dem verheiratet mit 95%							
geb. verheiratet mit geboren am 97%							
in verheiratet mit geboren in 100%							
Wohnung: Wohnung 98%							

Vordruck Nr. 304b (Weiblich univrch.) 9. 48 380 000

ISD Nr. 945 Staatsdruckerei Berlin Z'06

Baustellen = coming soon 😊

Automatisierte
Layouterkennung für
Tabellenstrukturen





Grundlagenworkshop (Woche vom 23.10-27.10).

In diesem 4-stündigen Workshop werden die Grundlagen und vielfältigen Möglichkeiten erläutert.

<https://terminplaner6.dfn.de/de/p/4ec3792a9c856eceddc9c4c8b55285bd-386134>

Transkribus Vertiefung + Transkribus Learn (Woche vom 13.11.-17.11. 2023).

Die erste Hälfte diese Workshops dient der Vertiefung und der Beantwortung von Fragen, die sich sicherlich seit dem 1. Workshop ergeben haben. In der zweiten Hälfte wird ein auf Transkribus aufbauendes Online-Learntool für die Begleitung handschriftenkundlicher Seminare vorgestellt.

<https://terminplaner6.dfn.de/de/p/a0d3f4ed0973fcf9c42ebc366bcdfc3a-386357>

Fragen?



Vielen Dank für Ihre Aufmerksamkeit

Fragen jetzt und auch gerne später:

Bruno Blüggel
blueggel@uni-greifswald.de

