UNIVERSITÄT GREIFSWALD
Wissen lockt. Seit 1456

Arbeiten mit Transkribus

Bruno Blüggel
blueggel@uni-greifswald.de

Inhalt

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

Kurze Historie

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?

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 – Transkribus-Web

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“

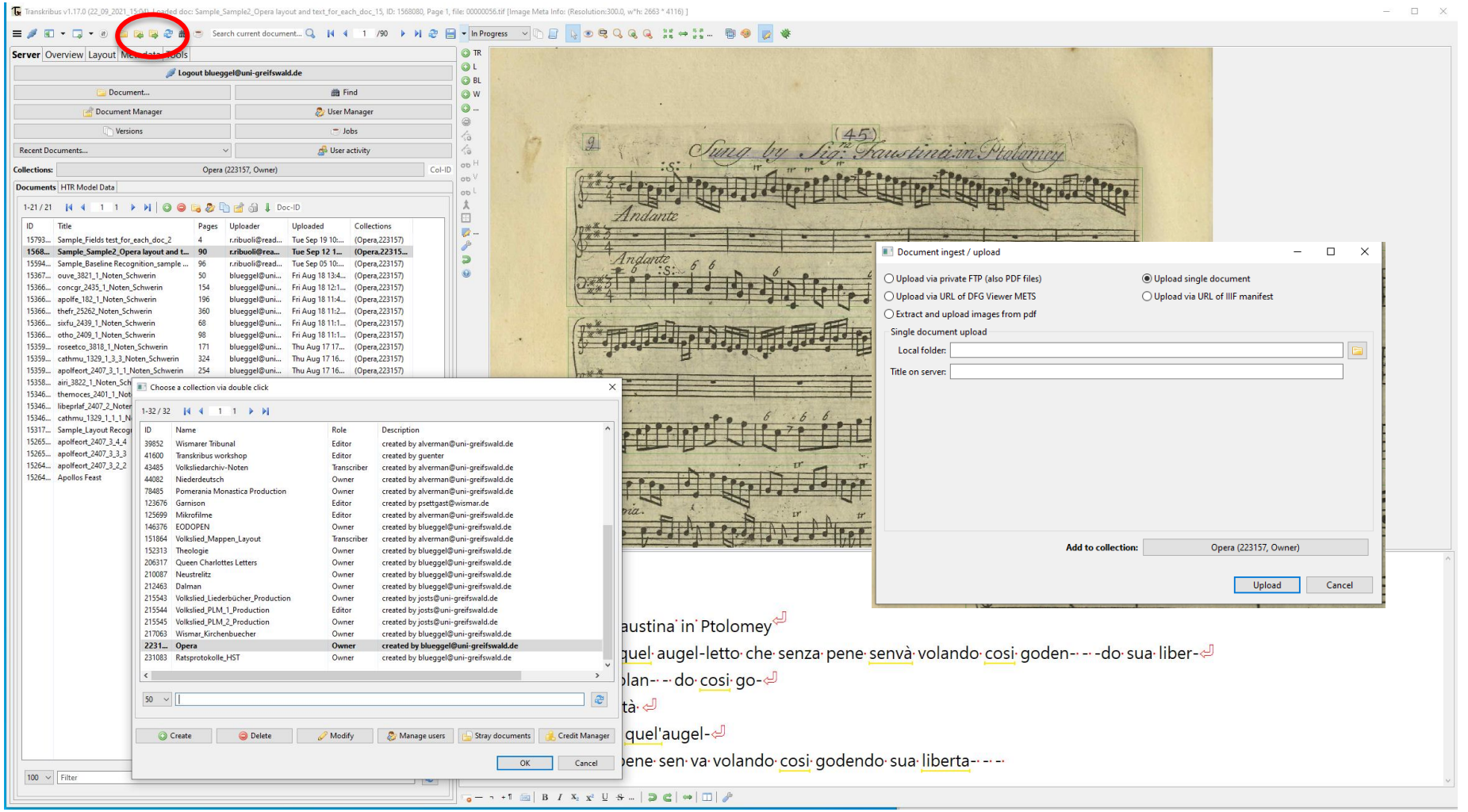
Transkribus-Client

Collections:

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

Hochladen von Images:

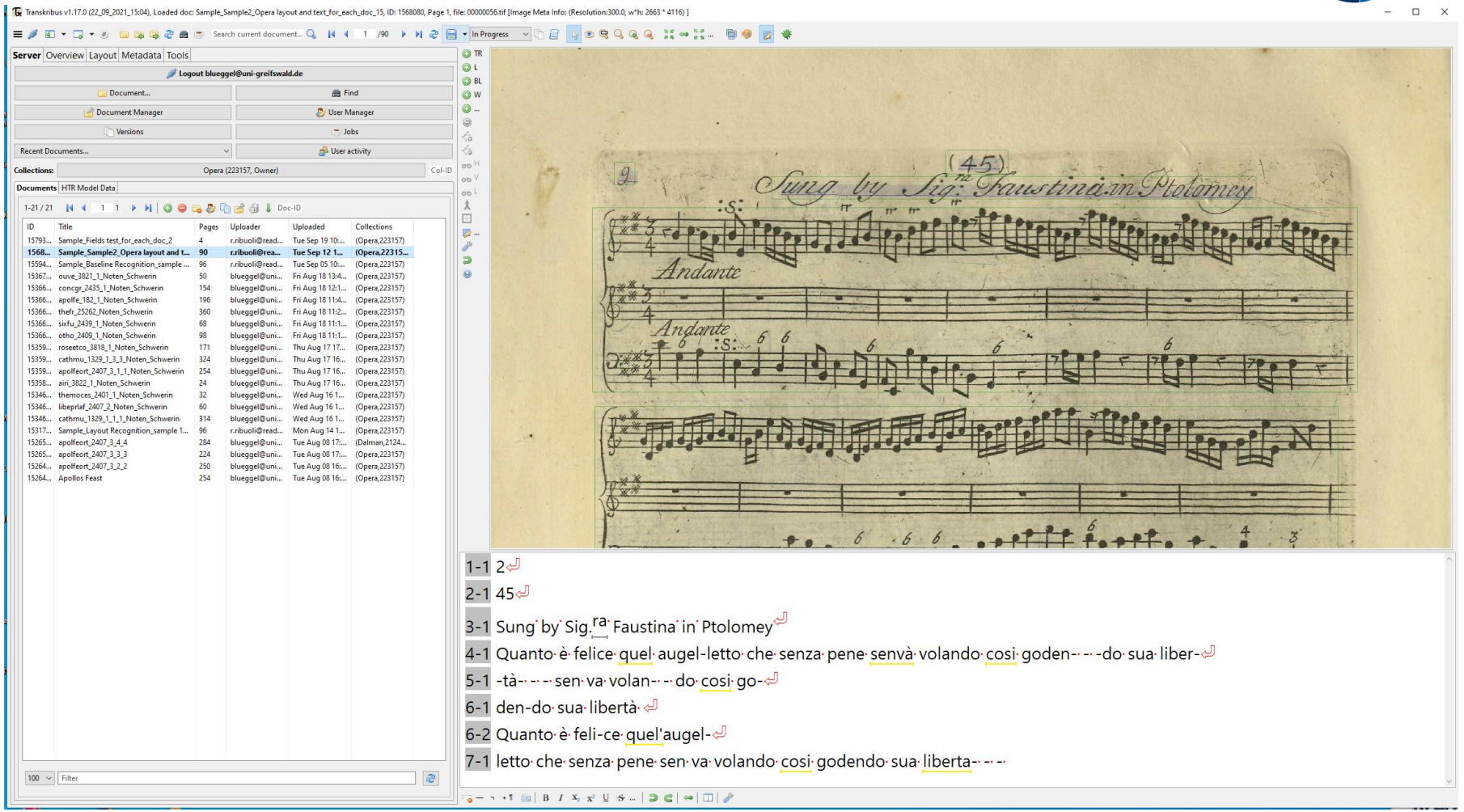
- JPG, PDF, ...
- per ftp client, lokale Verzeichnisse,



Transkribus-Client

Arbeitsoberfläche:

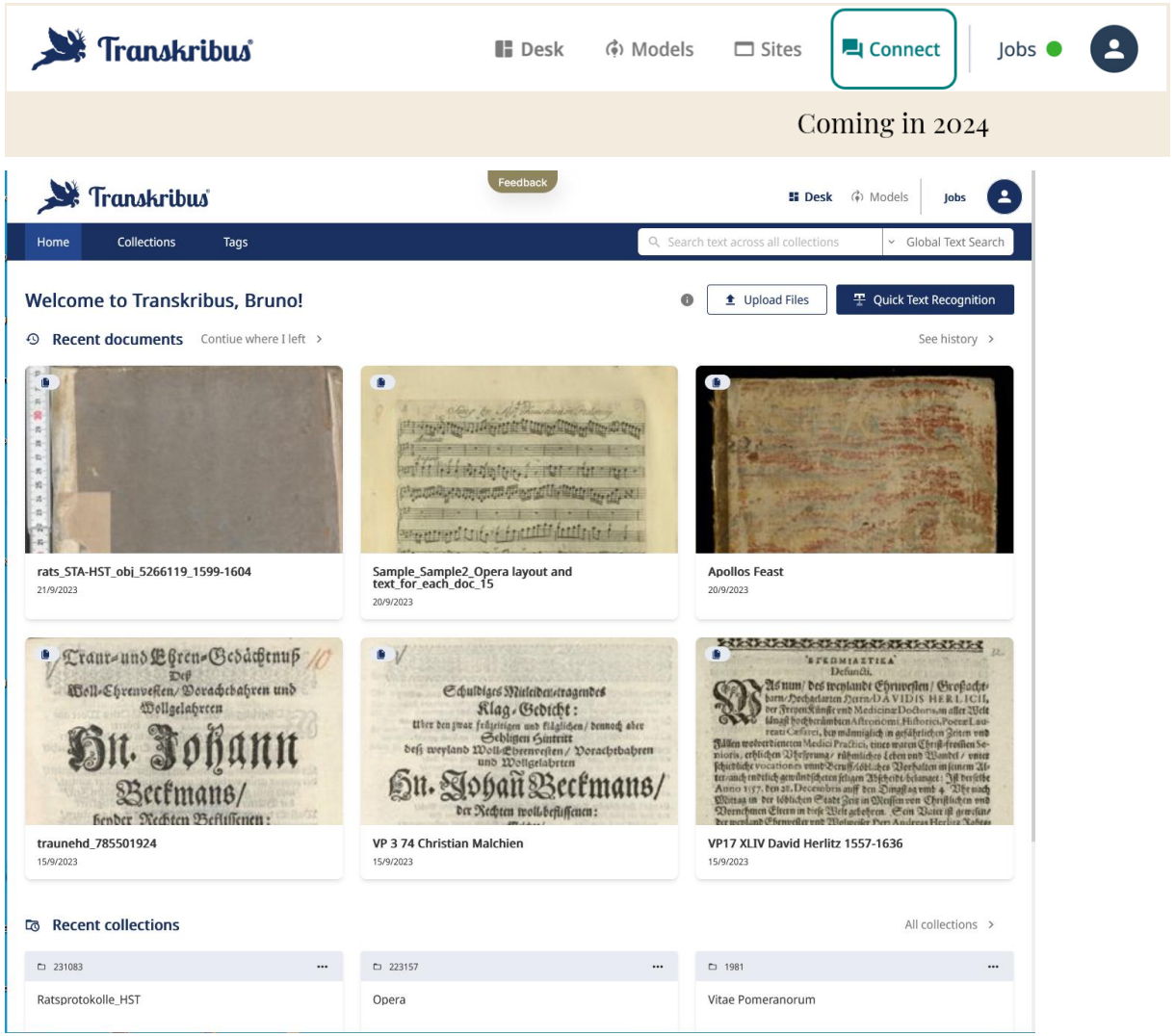
- Original +
- Editionsbereich
- Eine Vielzahl von Werkzeugen
- Aber etwas unübersichtlich
  → „Expertmode“



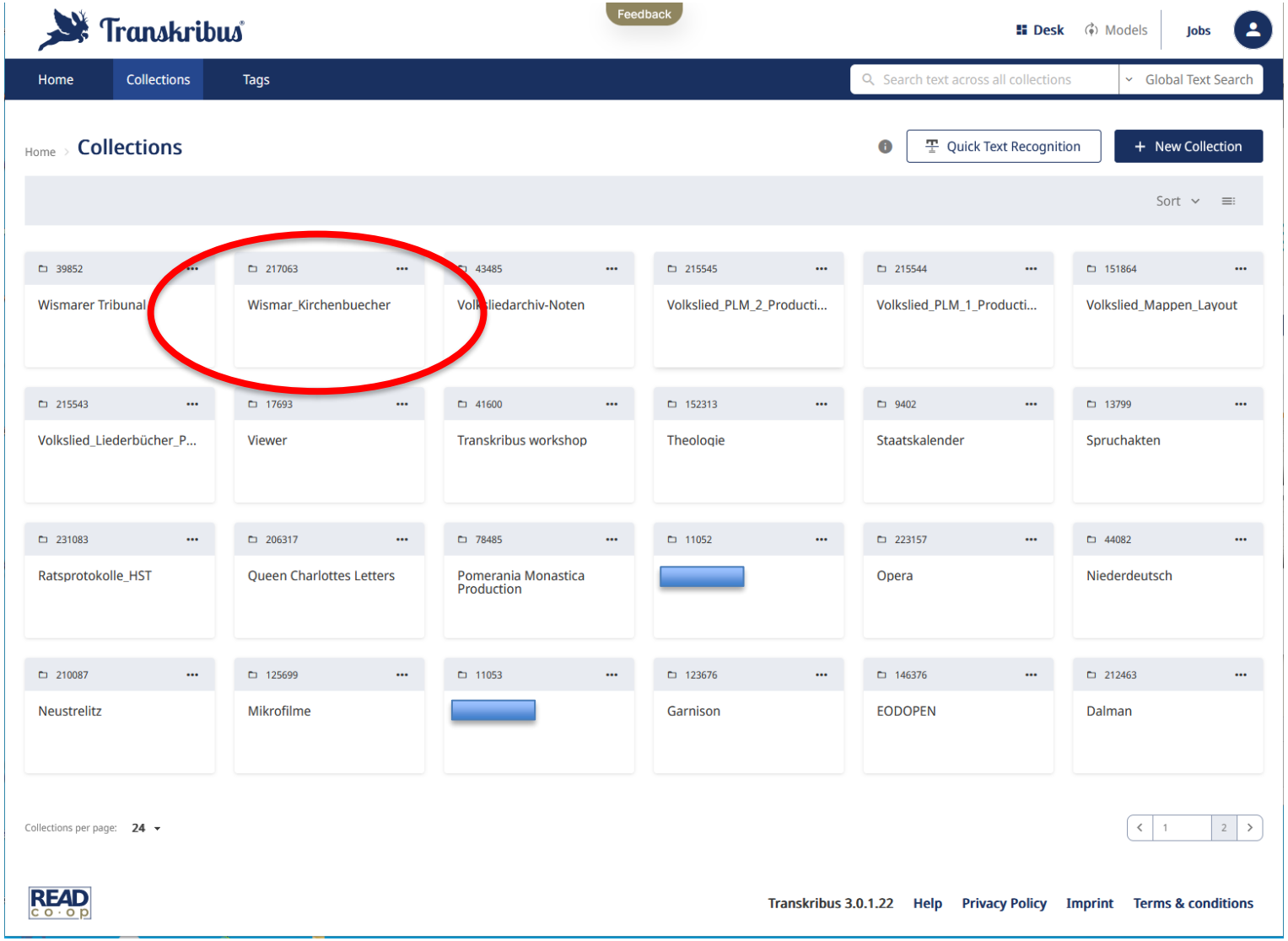
Transkribus-Web

Weiterentwicklung von „Transkribus Light“ zum neuen Interface

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



Transkribus-Web: Collections



Transkribus-Web: Collections

Farben symbolisieren „Bearbeitungsstände“

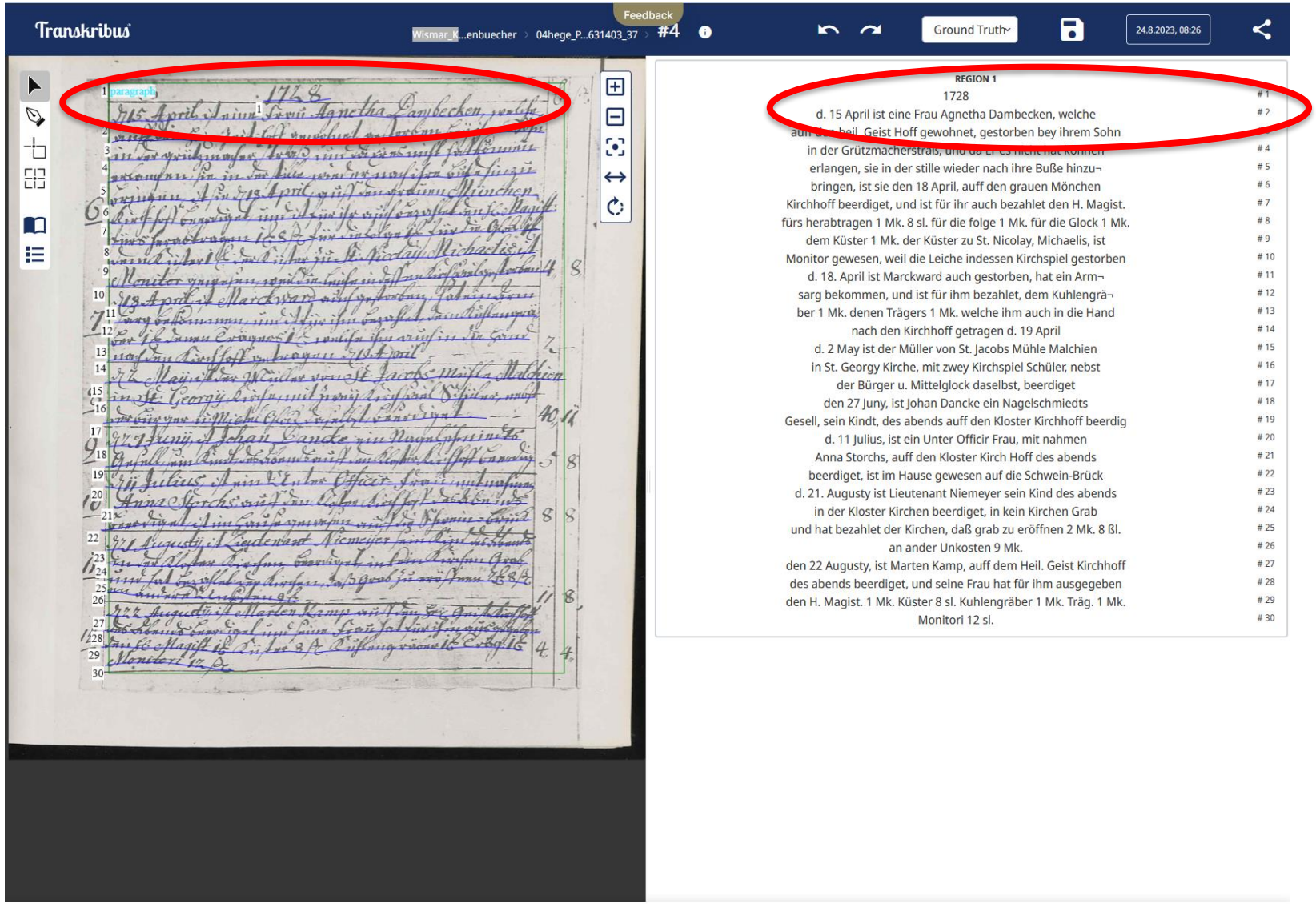
- Unbearbeitet
- In Arbeit
- Ground-Truth (100% korrekt)

Transkribus-Web: „Editor“

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

Vorgehen der KI:

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



Projektarbeit mit Transkribus

- 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

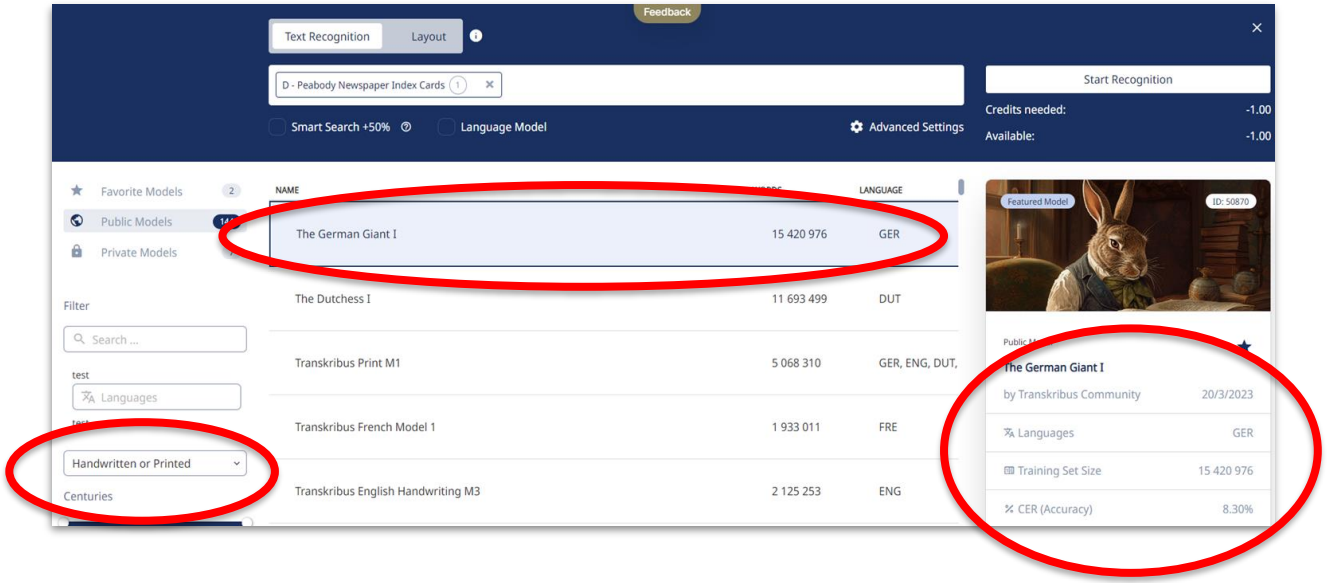
Farben symbolisieren „Bearbeitungsstände“

- 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 – 150 Public models

Modell auswählen

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



Training eines Modells

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?

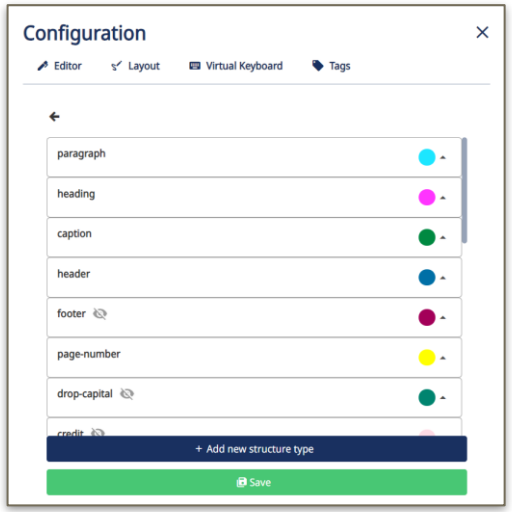
Public Models

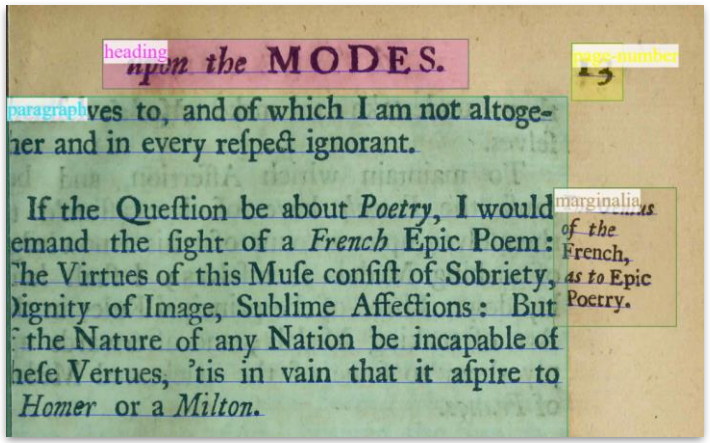
Die Performance eines Modells hängt ab:

- 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 Datenaustauch denken (z.B. Verlag)





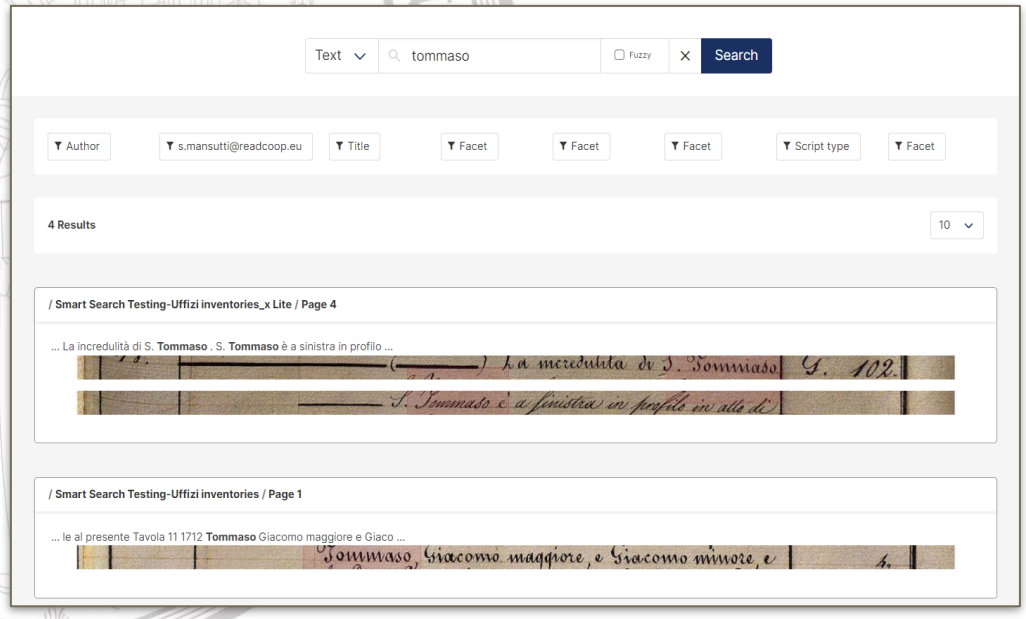
Suche

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



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)

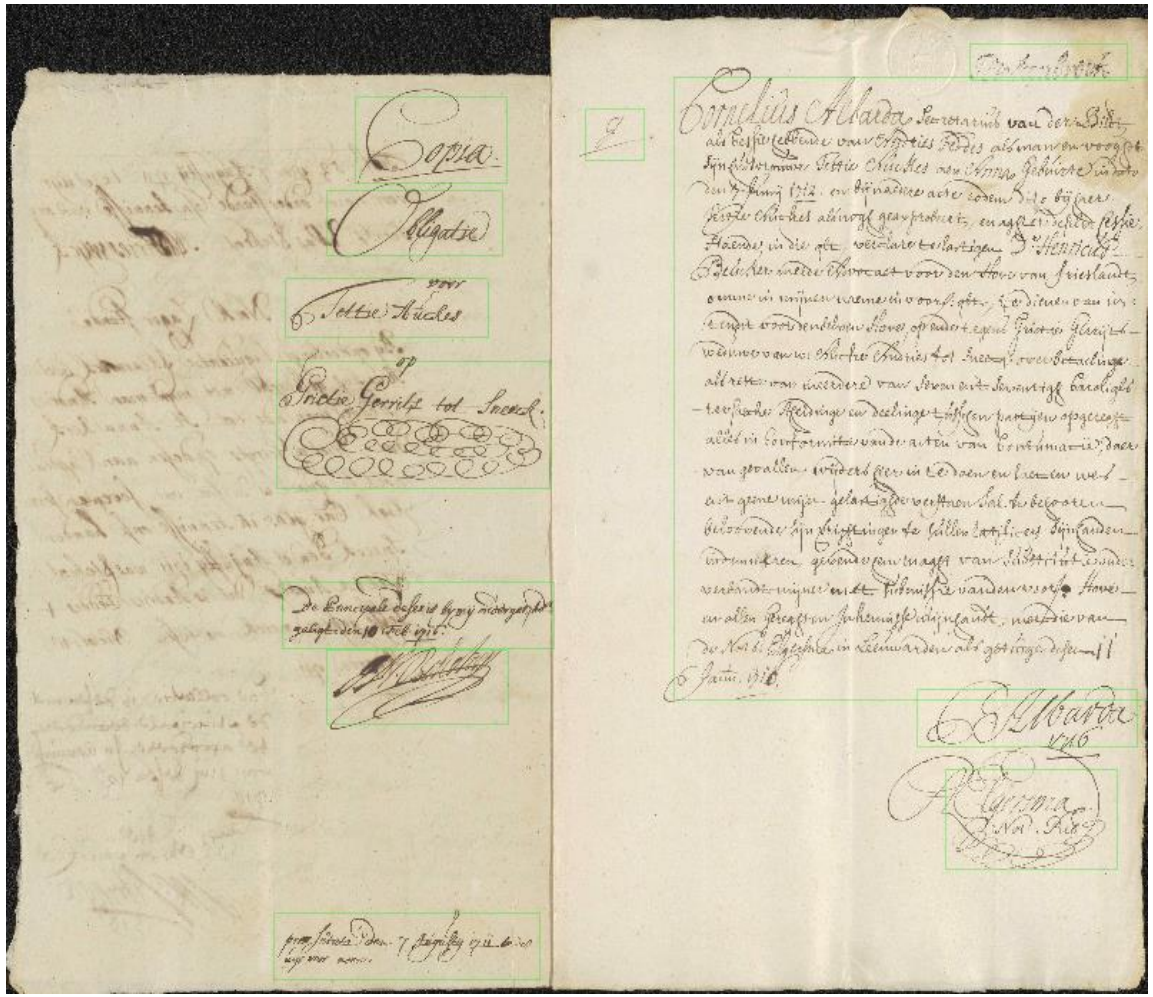
Baustellen = Beta Versionen ☺

Field Models (Beta Version)
Bisher ist beim Trainieren der Layouterkennung noch viel „händische“ Arbeit erforderlich
Automatisierte Layouterkennung für Zeitungssegmentierung

Baustellen = Betaversionen ☺

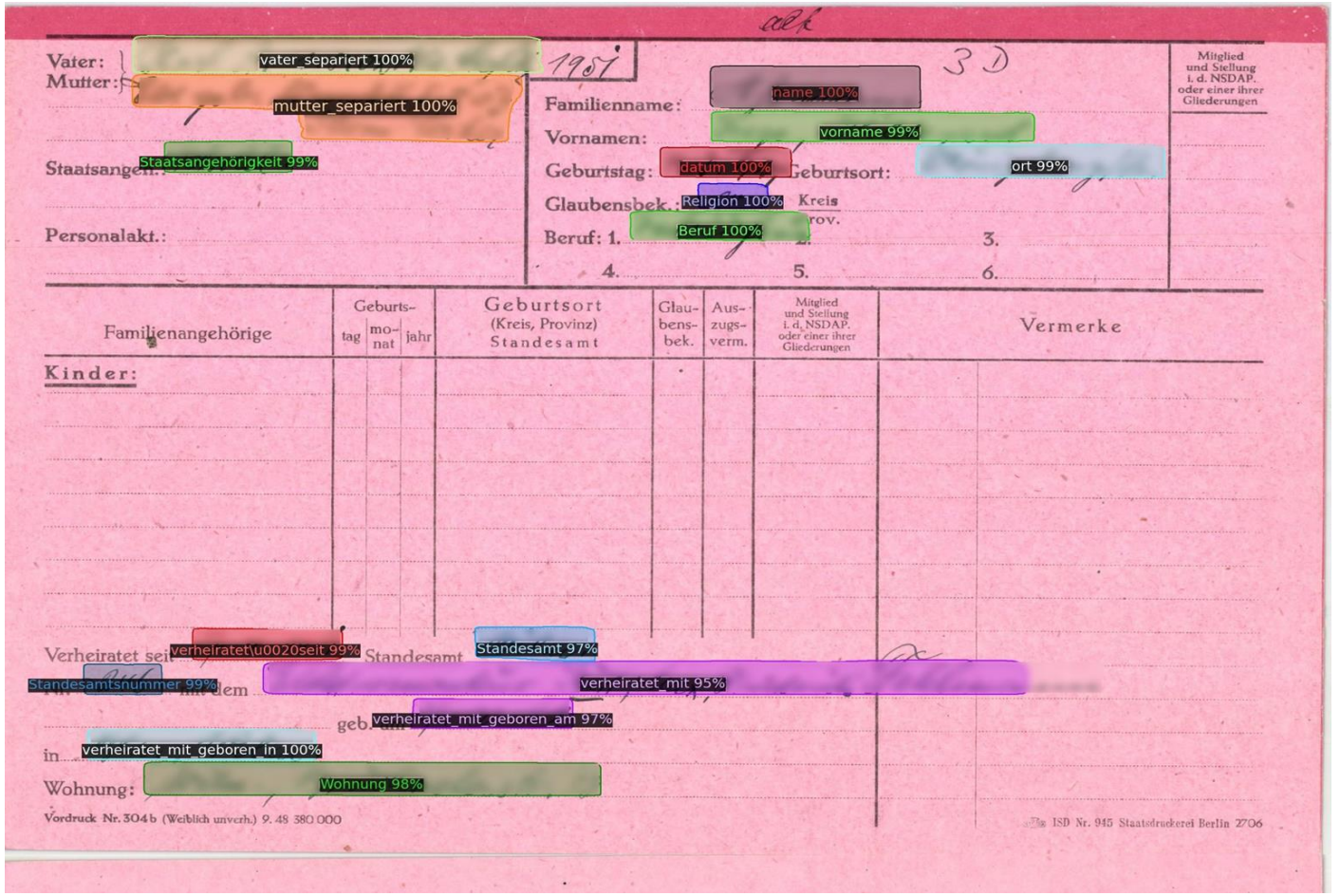
UNIVERSITÄT GREIFSWALD
Wissen lockt. Seit 1456

Automatisierte Layouterkennung für Text Regionen



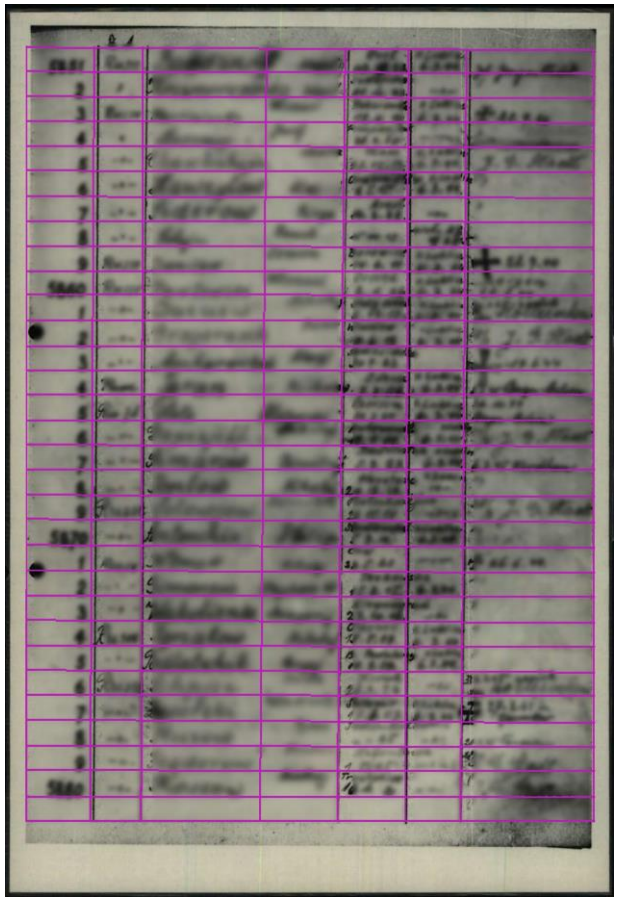
Baustellen = Betaversion ☺

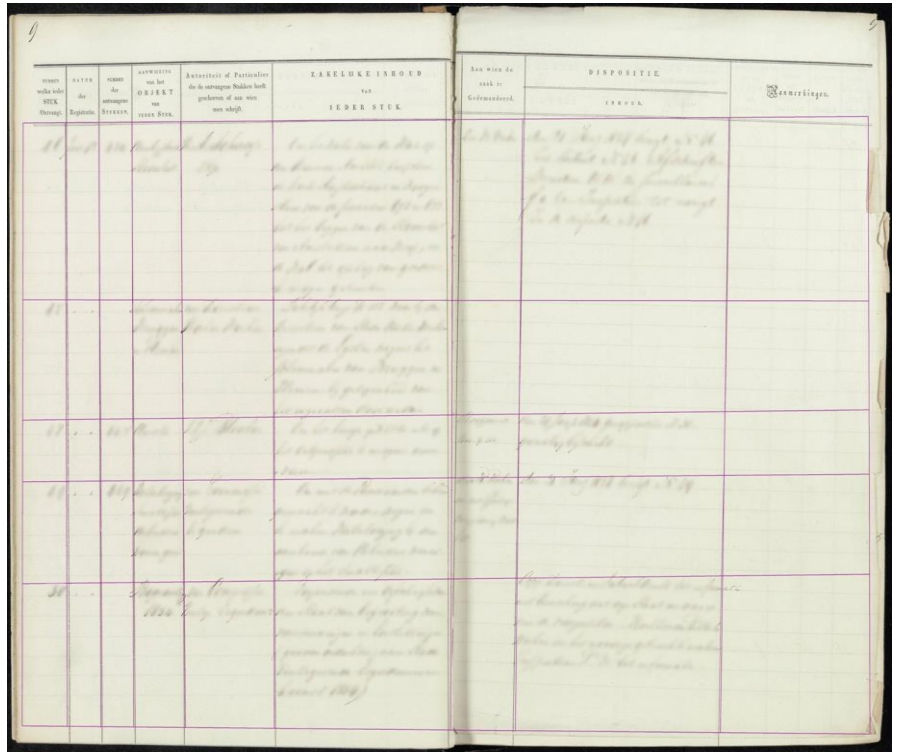
Automatisierte Layouterkennung
für Formaterkennung



Baustellen = coming soon ☺

Automatisierte Layouterkennung für Tabellenstrukturen





Trankribus Workshops

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 Häfte 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