



Hands-On-Lab

Volltexterschließung leicht gemacht:

Ein Crashkurs zur Transkriptionsplattform eScriptorium

Personen mit 2 337 812 074 **M.** versichert blieben, davon 552 246
Personen mit 1 847 622 742 **M.** bei den 35 Se im
Deutschen Reich. Der gesammte Zuwachs im Jahre 1877 stellt sich
auf 15 777 Personen mit 110 873 **M.**, d. h. 2,14 % der
Zahl der Versicherten und 4,98 % der Wachen,
summe. Die Berücksichtigung der Unfälle, welche die
Liquidation einiger Gesellschaften im Jahre 1877 zur Folge gehabt
hat vermindert sich derich den Zuwachs auf 8542 (1,16%) der Versicherten



Larissa Will, Jan Kamlah und Thomas Schmidt

Agenda

11:00 – 11:20 Uhr	Einführung <ul style="list-style-type: none">• OCR-Projekte der UB Mannheim• Transkriptionsplattformen Basics
11:20 – 12:20 Uhr	Hands-On <ul style="list-style-type: none">• Grundfunktionalitäten• (erweiterte) Layoutbearbeitung• Alignierung• Training• Ontologien
12:20 – 12:30 Uhr	Wrap Up <ul style="list-style-type: none">• Fragerunde

Übersicht

- ✓ 2018 – 2019: Optimierter Einsatz von OCR-Verfahren – Tesseract als Komponente im OCR-D-Workflow
- ✓ 2019 – 2022: OCR-BW – Kompetenzzentrum OCR der Universitätsbibliotheken Mannheim und Tübingen
- ✓ 2021 – 2023: OCR-D: Workflow für werkspezifisches Training auf Basis generischer Modelle mit OCR-D sowie Ground-Truth-Aufwertung
- ✓ 2021 – 2023: OCR-D: Integration von Kitodo und OCR-D zur produktiven Massendigitalisierung
- 2021 – 2026: BERD@NFDI: OCR-Services (TaskArea 3) für Business, Economic and Related Data

Services

- Individuelle Beratung von Forschenden und Studierenden von der Universität Mannheim, aber auch extern
- Durchführung einer offenen Sprechstunde einmal pro Monat
- Durchführung von Schulungen
- Durchführung von anderen Veranstaltungen wie z. B. Transcribathon
- Unterstützung bei Lehrveranstaltungen
- Hosting einer eScriptorium-Instanz

Einführung OCR

Wofür braucht man OCR?

- Durchsuchbare Volltexte generieren
- Komfortabler und niedrighschwelliger Zugang zu historischen Dokumenten ohne paläographische Kenntnisse
- Extraktion von Forschungsdaten



Wie funktioniert OCR?

- Textliche Bildinhalte → digitale Textformate
- OCR: „Optical Character Recognition“ (optische Zeichenerkennung)
- Begriff mittlerweile veraltet
- Neuronale Netzen erkennen nicht mehr Zeichen für Zeichen, sondern ganze Zeilen
- Texterkennung und OCR wird im deutschen Sprachraum oft synonym verwendet



Transkriptionsplattform Basics

Was ist eScriptorium?

- Transkriptionsplattform (OCR-Engine: [kraken](#))
- entwickelt an der Université Paris PSL
- Kostenfrei und Open-Source
- Alternative zu Transkribus
- Handgeschriebene und gedruckte Texte manuell oder automatisiert segmentieren und transkribieren
- Einfache Weitergabe trainierter Modelle
- Jeder kann eigene Instanz installieren
 - Für Windows, MacOS und Linux





Buchseite

span. Marokko, Angola. — Argentinien, Brasilien, Chile, Costa-Rica, Ecuador, Guatemala, Haiti, Kanada, Kuba, Mexiko, Paraguay, Peru, San Salvador, Uruguay, Venezuela, U. S. A.

Beteiligungen: Gesellschaft für Markt- u. Kühlhallen, Hamburg. — Blockeisfabrik Köln vorm. Gottfried Linde G.m.b.H., Köln. — Wasserstoff-, Sauerstoffwerke G.m.b.H., Schwarzenberg i. Sa. — Heylandt Gesellschaft für Apparatebau m. b. H., Berlin-Mariendorf. — Hydroxygen G. m. b. H., Wien. — Sauerstoff- u. Wasserstoffwerke A.-G., Luzern. — Dansk Ilt. & Brintfabrik A.-S., Kopenhagen. — Abello Oxigeno Linde S.-A., Barcelona. — Vereinigte Sauerstoff-Werke G. m. b. H., Berlin. — Elektro Zuur-en Waterstoffabriek, Amsterdam. — Linde-Riedinger Maschinenfabrik A.-G., Wien. — Marx & Traube G. m. b. H., Frankfurt a. M.

Besondere Angaben: Die Gesellschaft für Linde's Eismaschinen A.-G. konnte im Jahre 1929 auf ein erfolg-

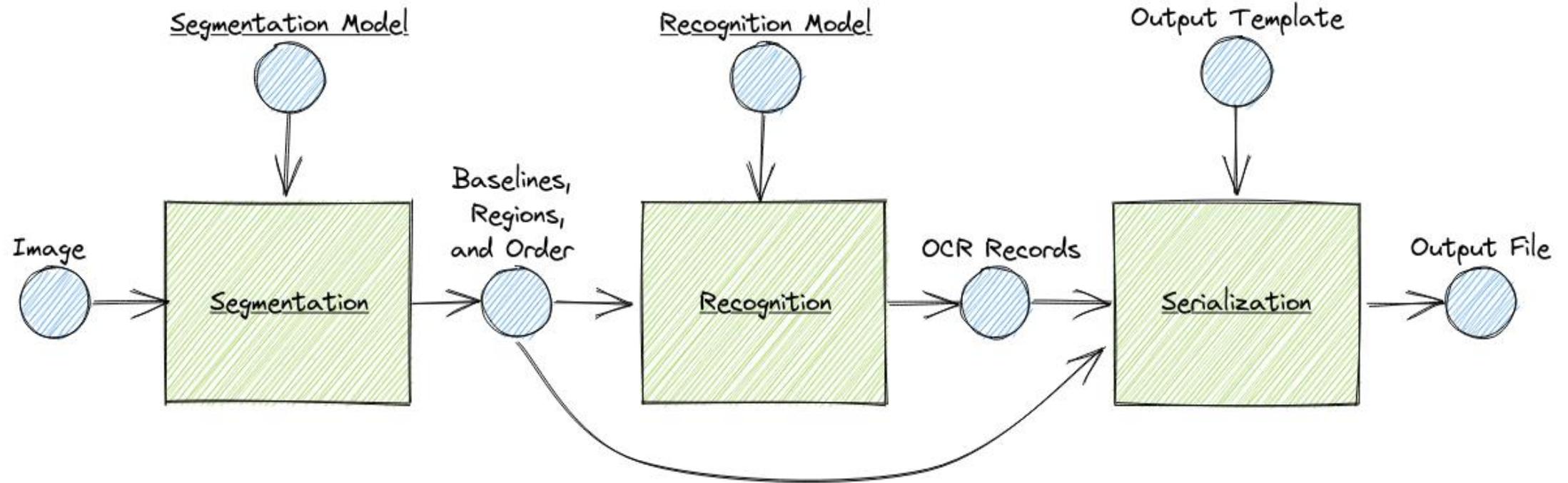
Transkription

span. Marokko, Angola. — Argentinien, Brasilien, Chile, Costa-Rica, Ecuador, Guatemala, Haiti, Kanada, Kuba, Mexiko, Paraguay, Peru, San Salvador, Uruguay, Venezuela, U. S. A.

Beteiligungen: Gesellschaft für Markt- u. Kühlhallen, Hamburg. — Blockeisfabrik Köln vorm. Gottfried Linde G.m.b.H., Köln. — Wasserstoff-, Sauerstoffwerke G.m.b.H., Schwarzenberg i. Sa. — Heylandt Gesellschaft für Apparatebau m. b. H., Berlin-Mariendorf. — Hydroxygen G. m. b. H., Wien. — Sauerstoff- u. Wasserstoffwerke A.-G., Luzern. — Dansk Ilt. & Brintfabrik A.-S., Kopenhagen. — Abello Oxigeno Linde S.-A., Barcelona. — Vereinigte Sauerstoff-Werke G. m. b. H., Berlin. — Elektro Zuur-en Waterstoffabriek, Amsterdam. — Linde-Riedinger Maschinenfabrik A.-G., Wien. — Marx & Traube G. m. b. H., Frankfurt a. M.

Besondere Angaben: Die Gesellschaft für Linde's Eismaschinen A.-G. konnte im Jahre 1929 auf ein erfolg-

Einführung OCR



Bildcredit: <https://kraken.re/main/index.html>

Hands-On



Gehen Sie auf: <https://ocr-bw.bib.uni-mannheim.de/escriptorium/>



Melden Sie sich mit Ihren Zugangsdaten an:

→ Benutzername: **workshop2** | PW: **bibliocon2024!**



Wir gehen zusammen Schritt für Schritt durch eScriptorium!

Workflow-Schritt: Dokument anlegen

eScriptorium Home Kontakt Suche in Amtsschrifttum Meine Projekte Meine Modelle Hallo Thomas

Amtsschrifttum 1800 Dokumente Bearbeiten Berichte Neues Dokument anlegen

Auswählen Vorschau Name Besitzer Zuletzt geändert Anzahl der Bilder Schlagwörter Aktionen

eScriptorium Home Kontakt AI Suche in 1800 VI bis 192 Meine Projekte Meine Modelle Hallo Thomas

Beschreibung **Ontologie** Bilder Bearbeiten Modelle

Document created successfully!

1800 VI bis 1920 IX

Links nach rechts

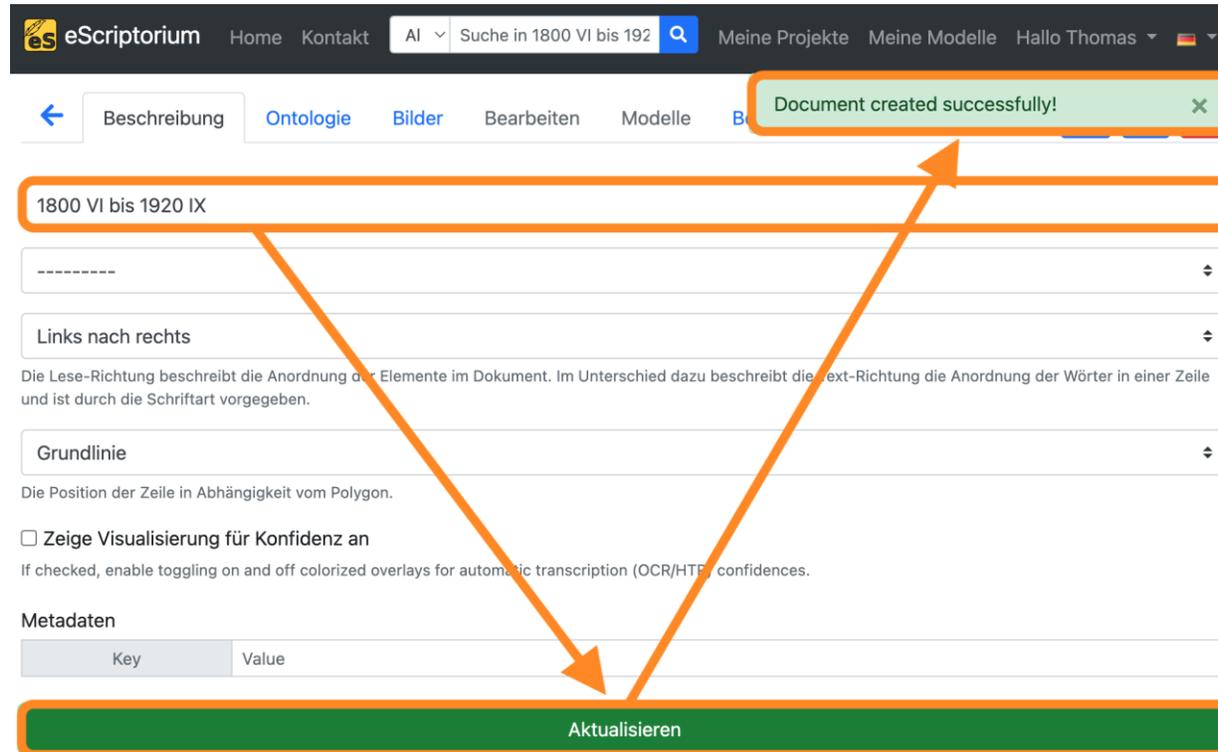
Grundlinie

Zeige Visualisierung für Konfidenz an

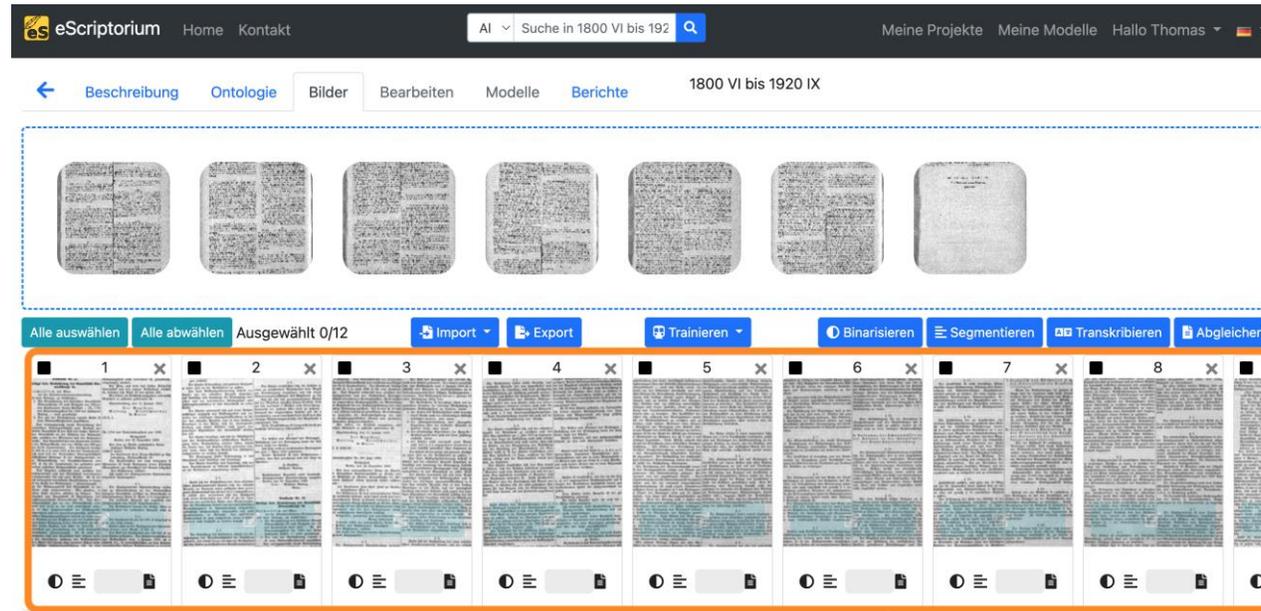
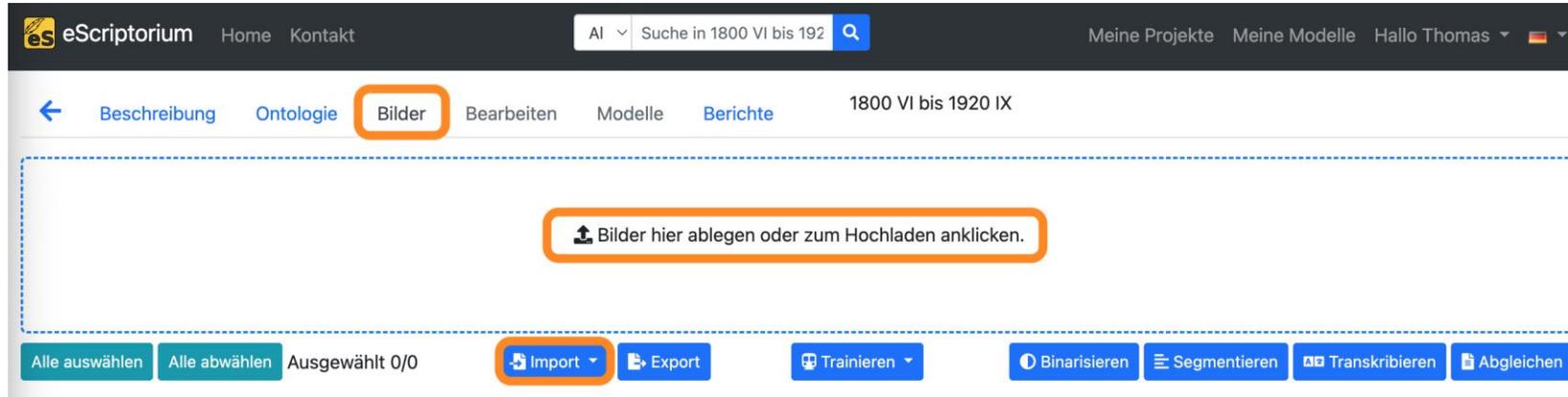
Metadaten

Key	Value
-----	-------

Aktualisieren



Workflow-Schritt: Bilder hochladen

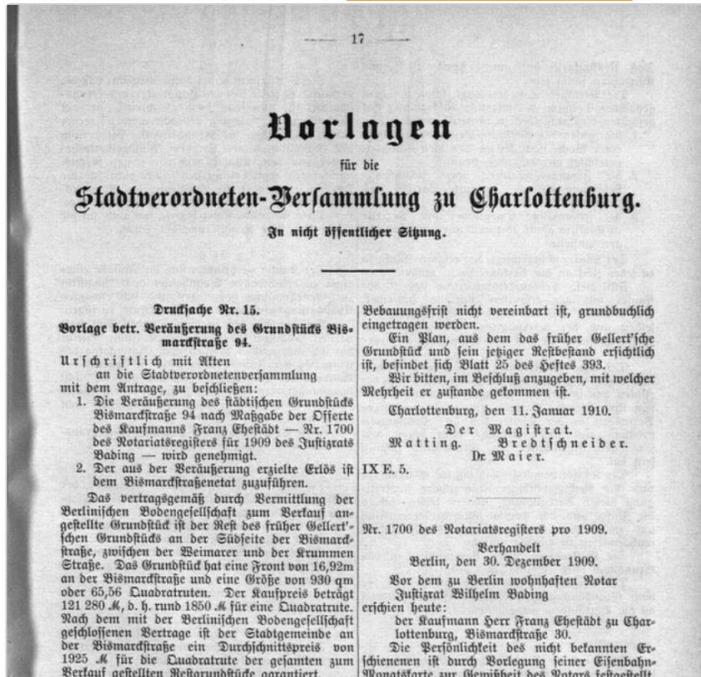


Theorie: Ergebnis der Segmentierung

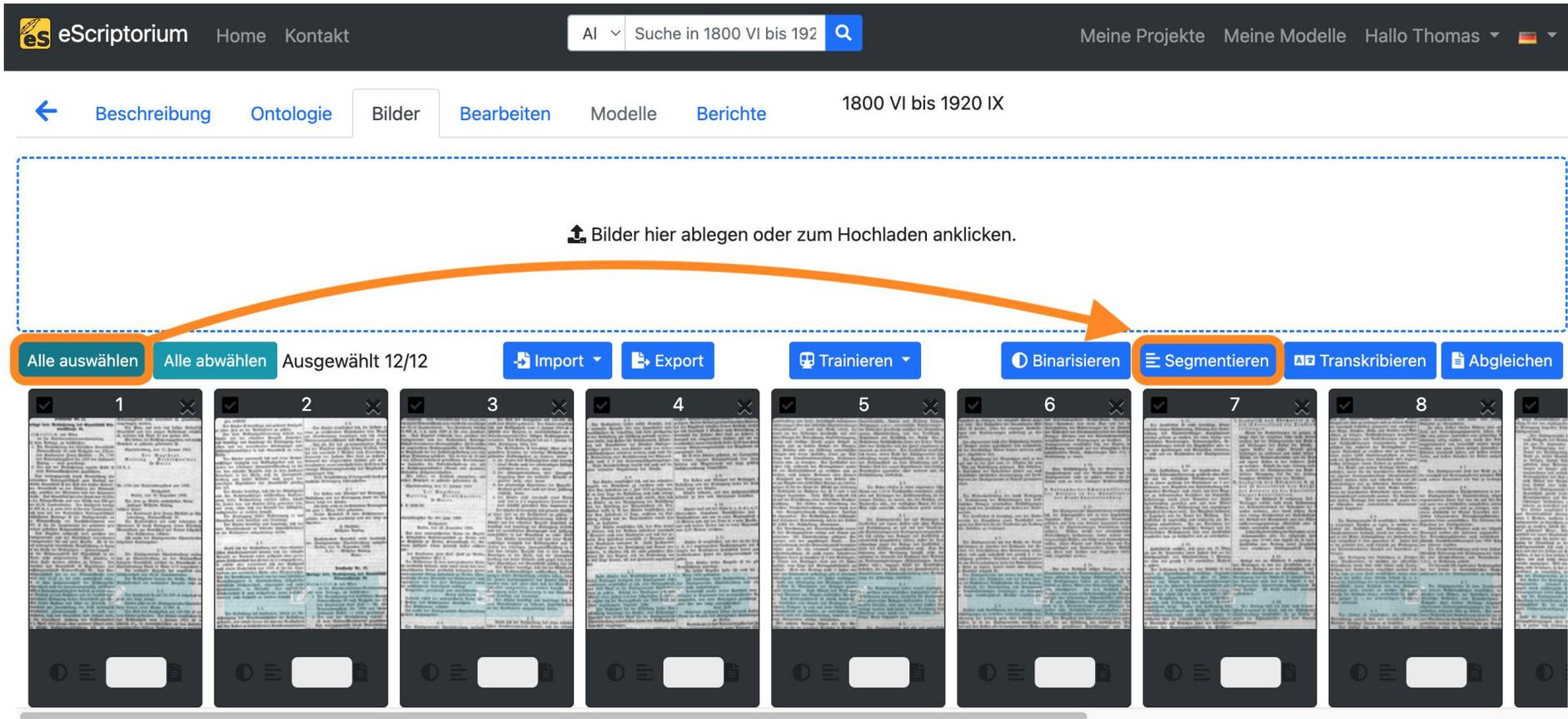


Buchseite

Layoutsegmentierung



Workflow-Schritt: Bild segmentieren



The screenshot shows the eScriptorium interface. At the top, there is a navigation bar with the eScriptorium logo, 'Home', and 'Kontakt'. A search bar contains 'AI' and 'Suche in 1800 VI bis 192'. On the right, there are links for 'Meine Projekte', 'Meine Modelle', and a user profile 'Hallo Thomas' with a German flag. Below the navigation bar, a breadcrumb trail shows 'Beschreibung', 'Ontologie', 'Bilder', 'Bearbeiten', 'Modelle', and 'Berichte'. The current view is '1800 VI bis 1920 IX'. A large dashed blue box contains the text 'Bilder hier ablegen oder zum Hochladen anklicken.' with an upload icon. Below this box is a toolbar with buttons: 'Alle auswählen' (highlighted with an orange box), 'Alle abwählen', 'Ausgewählt 12/12', 'Import', 'Export', 'Trainieren', 'Binarisieren', 'Segmentieren' (highlighted with an orange box), 'Transkribieren', and 'Abgleichen'. An orange arrow points from the 'Alle auswählen' button to the 'Segmentieren' button. Below the toolbar is a grid of 12 image thumbnails, numbered 1 to 8, each with a checkmark and a close button. The thumbnails show scanned pages of text with some areas highlighted in blue.

Workflow-Schritt: Bild segmentieren

12 ausgewählte(s) Bild(er). ?

Achtung - eine Re-Segmentierung wird alle bestehenden Transkriptionen löschen! ✕

Wählen Sie ein Modell aus

default (bla.mlmodel) ▾

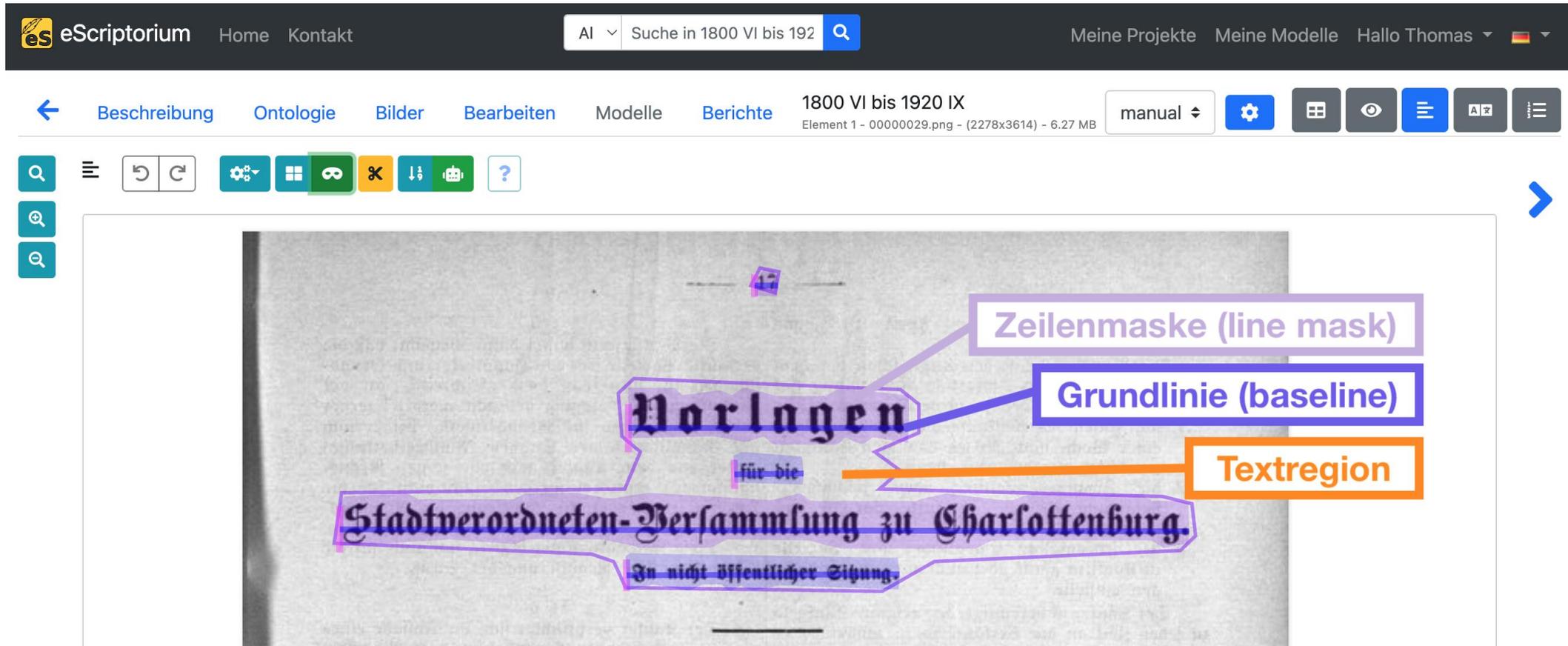
Zeilen und Bereiche ▾

Horizontal v. l. n. r. ▾

Override
Wenn diese Option aktiviert ist, werden vorhandene Segmentierungen **und damit verbundene Transkriptionen** zuerst gelöscht!

Close Segmentieren

Theorie: Ergebnis der Segmentierung



The screenshot shows the eScriptorium web interface. At the top, there is a navigation bar with the eScriptorium logo, links for Home and Kontakt, a search bar containing 'Suche in 1800 VI bis 192', and user information 'Meine Projekte', 'Meine Modelle', and 'Hallo Thomas'. Below this is a secondary navigation bar with tabs for 'Beschreibung', 'Ontologie', 'Bilder', 'Bearbeiten', 'Modelle', and 'Berichte'. The current view is 'Berichte' for the document '1800 VI bis 1920 IX'. A toolbar on the left contains various icons for search, zoom, and editing. The main content area displays a document image with several text segments highlighted in purple. Three callout boxes with lines pointing to these segments are present: a purple box labeled 'Zeilenmaske (line mask)' pointing to the top line of the first segment, a blue box labeled 'Grundlinie (baseline)' pointing to the bottom edge of the first segment, and an orange box labeled 'Textregion' pointing to the entire first segment. The text in the image is in a Gothic script and reads: 'Vorlagen für die Stadtverordneten-Versammlung zu Charlottenburg. In nicht öffentlicher Sitzung.'

Layoutbearbeitung

- **Veränderung der Polygonzüge** (Regionen und Zeilenmasken):
 - Doppelklick fügt dem Polygonzug **neuen Punkt** hinzu
 - Shift + linke Maustaste wählt **mehrere Punkte eines Polygonzugs** aus
 - Strg + linker Maustaste verschiebt die Auswahl
 - „c“-Taste **aktiviert das Cutting-Tool**:
 - mit diesem Tool lassen sich Regionen / Textzeilen trennen
 - Verschiedene Baselines / Textzeilen können mit Shift + linker Maustaste ausgewählt und über die Schaltfläche  verbunden werden
 - Hier ist auch Stapelverarbeitung möglich (bspw. gleichzeitige Verkürzung mehrerer Baselines)

Datensätze + Modelle importieren und bereitstellen

- **Bestehende** Transkription importieren
 - XML-Format: Upload von ALTO und PAGE-XML
 - Plain-Text: **Alignment (Abgleich)**
 - *Line length match threshold:* 0,4
 - *N-Gramm:* 10
 - *Beam:* 100
- Erstellte Transkriptionen und Modelle **bereitstellen**
 - Export als PAGE / ALTO (Transkriptionen)
 - Upload auf GitHub und Zenodo (OCR/HTR Community https://zenodo.org/communities/ocr_models/),
 - HTR-United <https://htr-united.github.io/>

Neuronale Netze müssen trainiert werden!

- **garbage in → garbage out**
- **Iterative Trainingszyklen** nutzen:
→ Handvoll Seiten transkribieren → Training → Test → Trainingsset erweitern
- Auf **repräsentative Verteilung** des Zeichensatzes achten
- **Was geht (derzeit) nicht in eS?**
 - Tabellen / kompliziertes Layout
 - Große Seiten / Landkarten
 - Bearbeitung der Zeilenmasken sehr umständlich

Empfehlungen

- **Segmentierung**
 - Min. 10-20 Seiten des gleichen Seitentyps für das erste Training
 - Wahl eines Basismodells das möglichst die gewünschten TextRegionen-Typen bereits abdeckt
 - Baselines und Textlinienpolygone können im Trainingsset etwas größer bemessen werden
 - TextRegion des gleichen Typs sollten sich möglichst nicht überlappen
- **Transkription**
 - Min. 5-10 Seiten (möglichst hohe Abdeckung des Zeichenvorrats)
 - Wahl eines zum Schriftsystem passenden Basismodells mit gleichem Zeichenvorrat
 - Zeilen mit seltenen Sonderzeichen und Zeichen mit geringer Erkennungsraten bevorzugen
 - Zeichenvariation berücksichtigen

Verwendung von Ontologien

- **Bereichstypen:**
 - frei bestimmbare Label für Regionen (bspw. `Paragraph`, `TextRegion`, `Illustration` ...)
- **Zeilentypen:**
 - frei bestimmbare Label für Textzeilen (bspw. `heading`, `subheading`, `footnote` ...)
- Mit `Shift` + linker Maustaste können mehrere Regionen / Textzeilen gelabelt werden
- „Annotationskomponenten“ werden derzeit nicht exportiert!

Vielen Dank für Ihre Aufmerksamkeit!

Larissa Will (larissa.will@uni-mannheim.de)

Jan Kamlah (jan.kamlah@uni-mannheim.de)

Thomas Schmidt (thomas.schmidt@uni-mannheim.de)

Nächster Termin:

Offene OCR-Sprechstunde, immer am zweiten Donnerstag im Monat von 15 – 16 Uhr

- <https://ocr-bw.bib.uni-mannheim.de/>
- Nächster Termin: Donnerstag, 13. Juni 2024