FDZ FORSCHUNGSDATENZENTRUM UB MANNHEIM

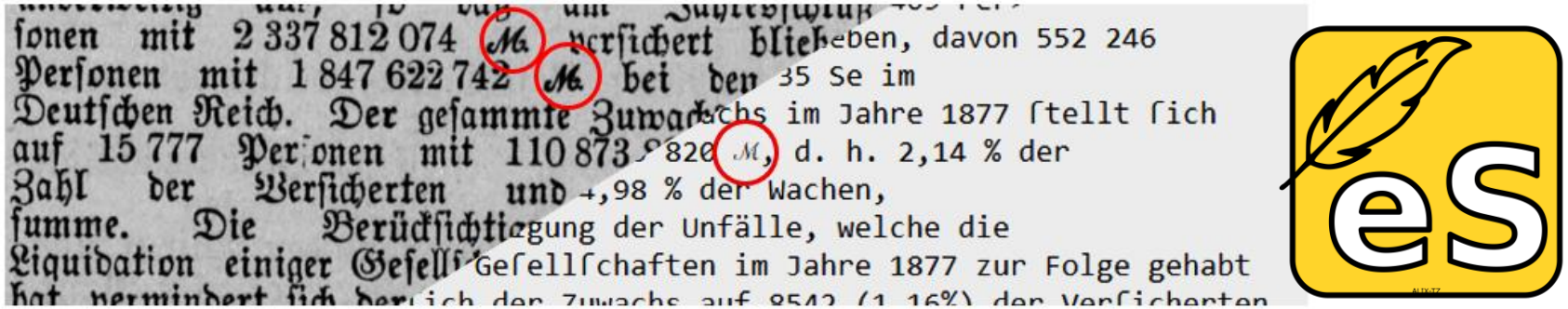
BERD @NFDI

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

Hands-On-Lab

Volltexterschließung leicht gemacht:

Ein Crashkurs zur Transkriptionsplattform eScriptorium



Larissa Will, Jan Kamlah und Thomas Schmidt

Agenda

11:00 – 11:20 Uhr	**Einführung** • OCR-Projekte der UB Mannheim • Transkriptionsplattformen Basics
11:20 – 12:20 Uhr	**Hands-On** • Grundfunktionalitäten • (erweiterte) Layoutbearbeitung • Alignierung • Training • Ontologien
12:20 – 12:30 Uhr	**Wrap Up** • Fragerunde

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

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

UB
UNIVERSITÄTSBIBLIOTHEK
MANNHEIM

Einführung OCR

BiblioCon 2024 | Larissa Will, Jan Kamlah, Thomas Schmidt | 06.06.2024

Wofür braucht man OCR?

- Durchsuchbare Volltexte generieren
- Komfortabler und niedrigschwelliger 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

UB
UNIVERSITÄTSBIBLIOTHEK
MANNHEIM

Transkriptionsplattform Basics

BiblioCon 2024 | Larissa Will, Jan Kamlah, Thomas Schmidt | 06.06.2024

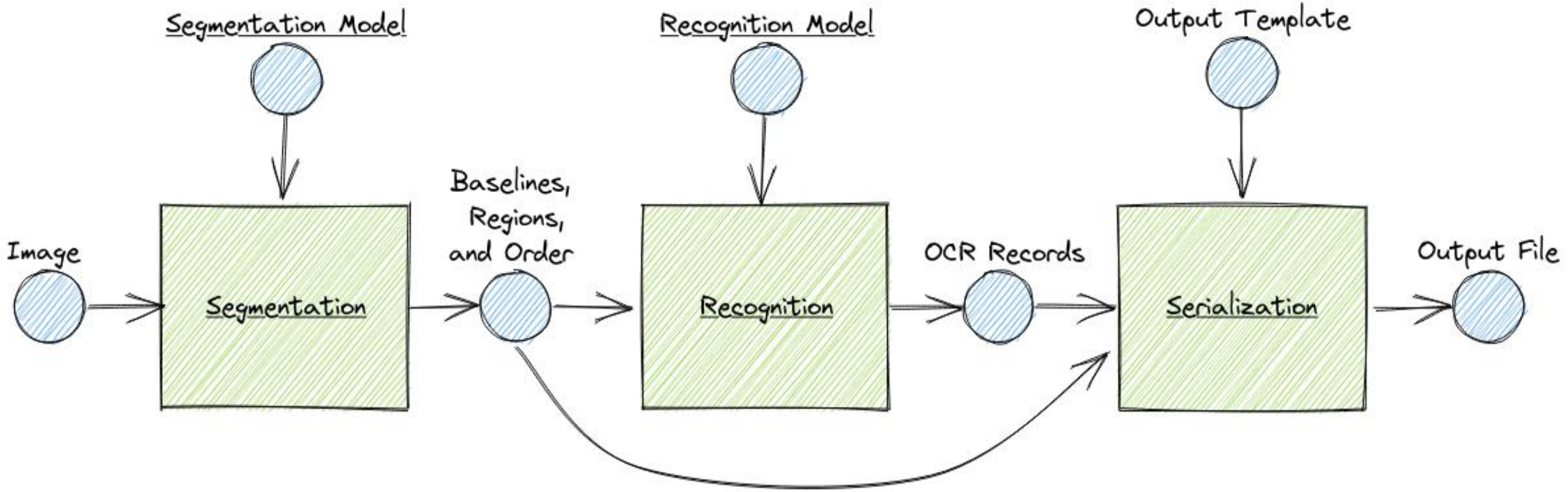
UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

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

Einführung OCR

Einführung OCR



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

UB
UNIVERSITÄTSBIBLIOTHEK
MANNHEIM

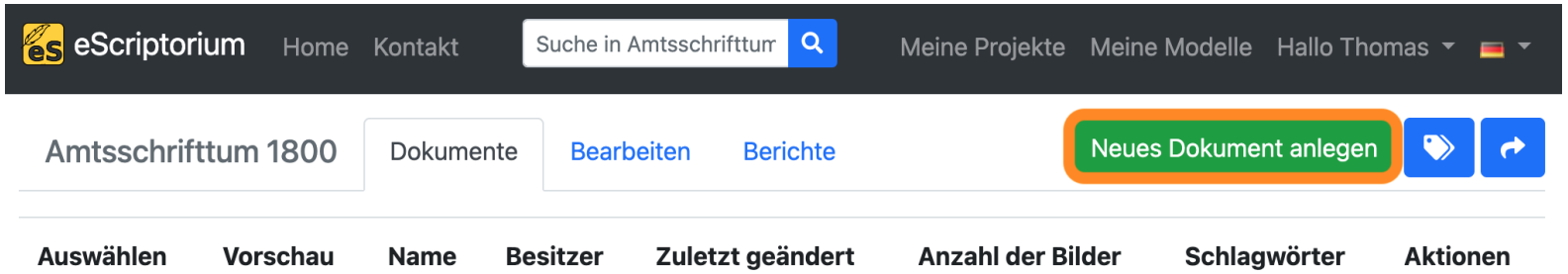
Hands-On

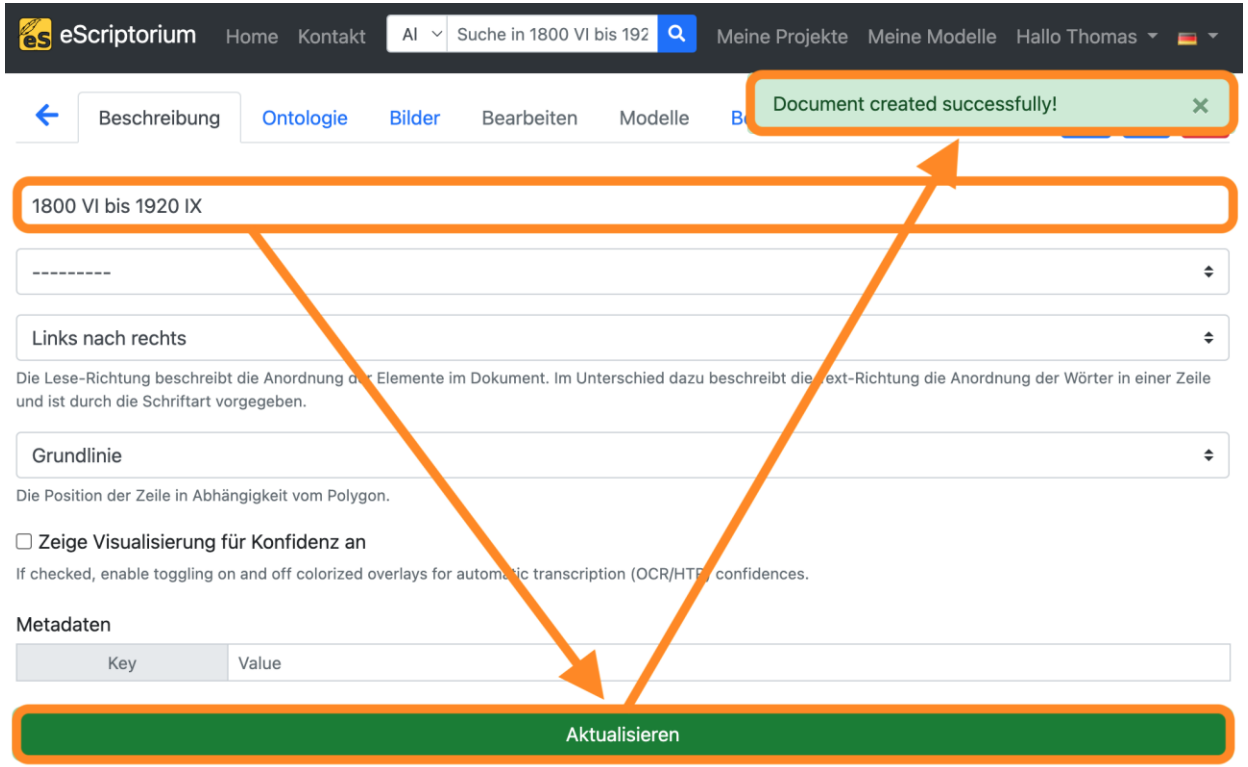
Gehen Sie auf: https://ocr-bw.bib.uni-mannheim.de/escriptorium/

Melden Sie sich mit Ihren Zugangsdaten an:
→ Benutzername: **workshop2** | PW: **bibliocon2024!**

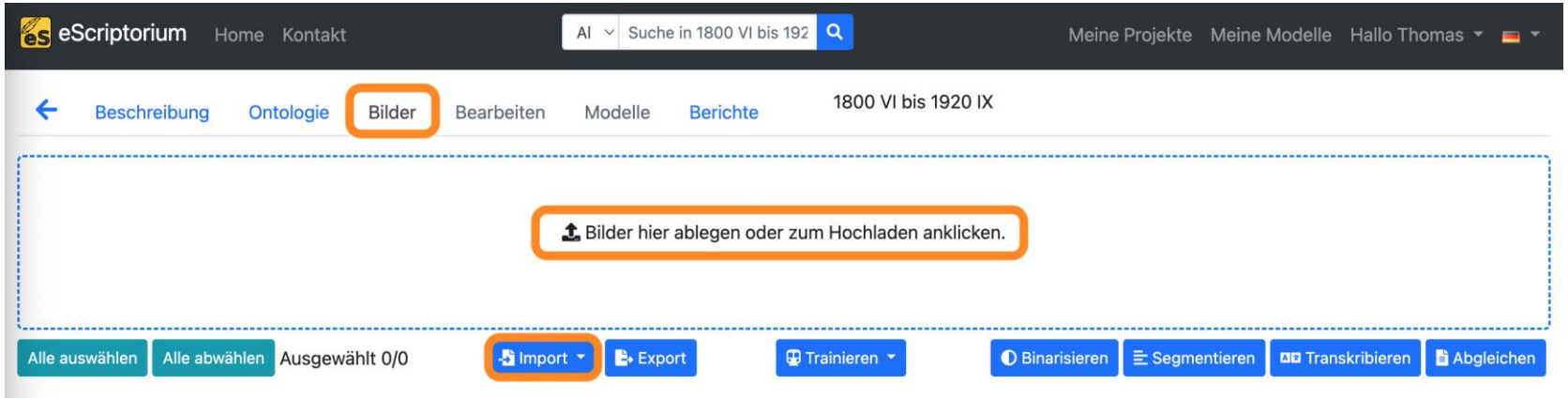
Wir gehen zusammen Schritt für Schritt durch eScriptorium!

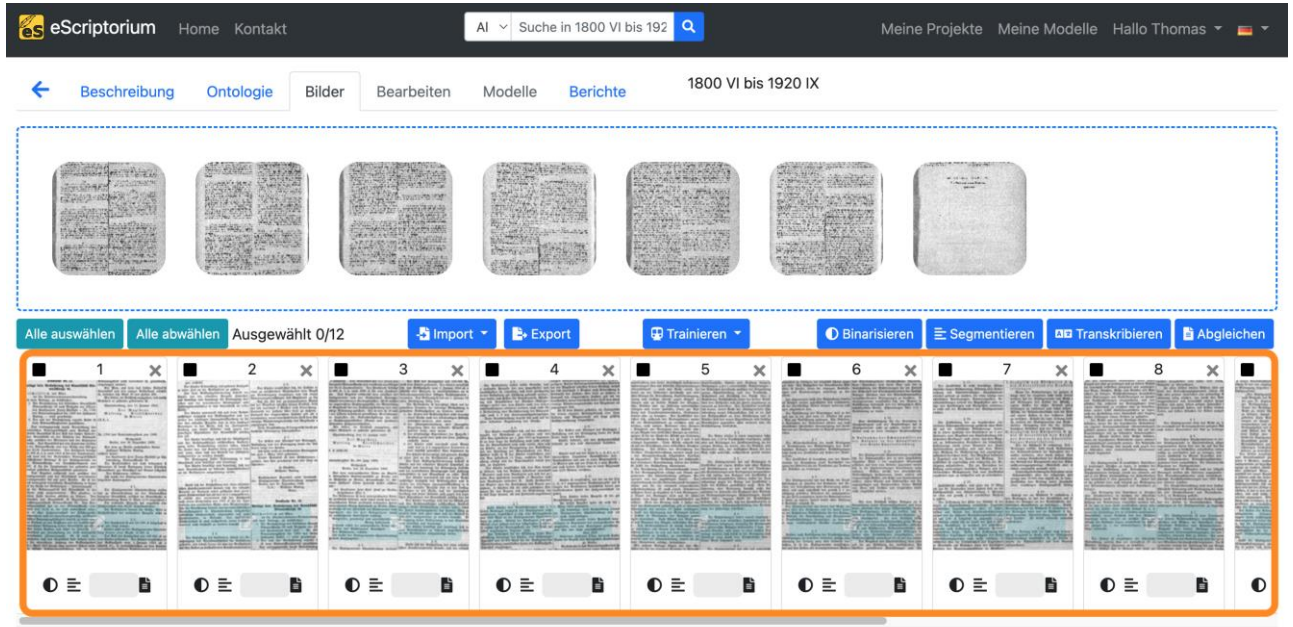
Workflow-Schritt: Dokument anlegen



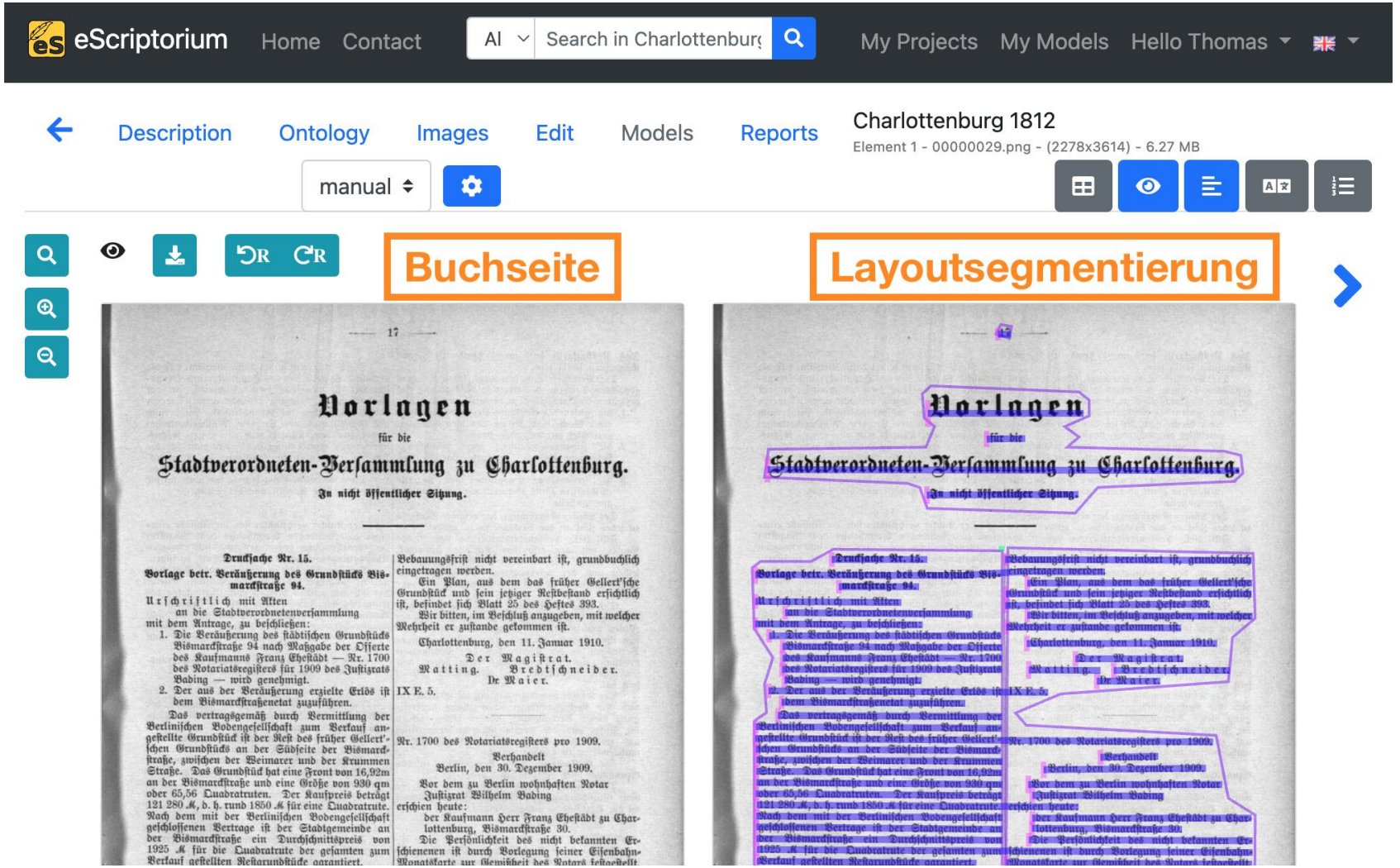


Workflow-Schritt: Bilder hochladen



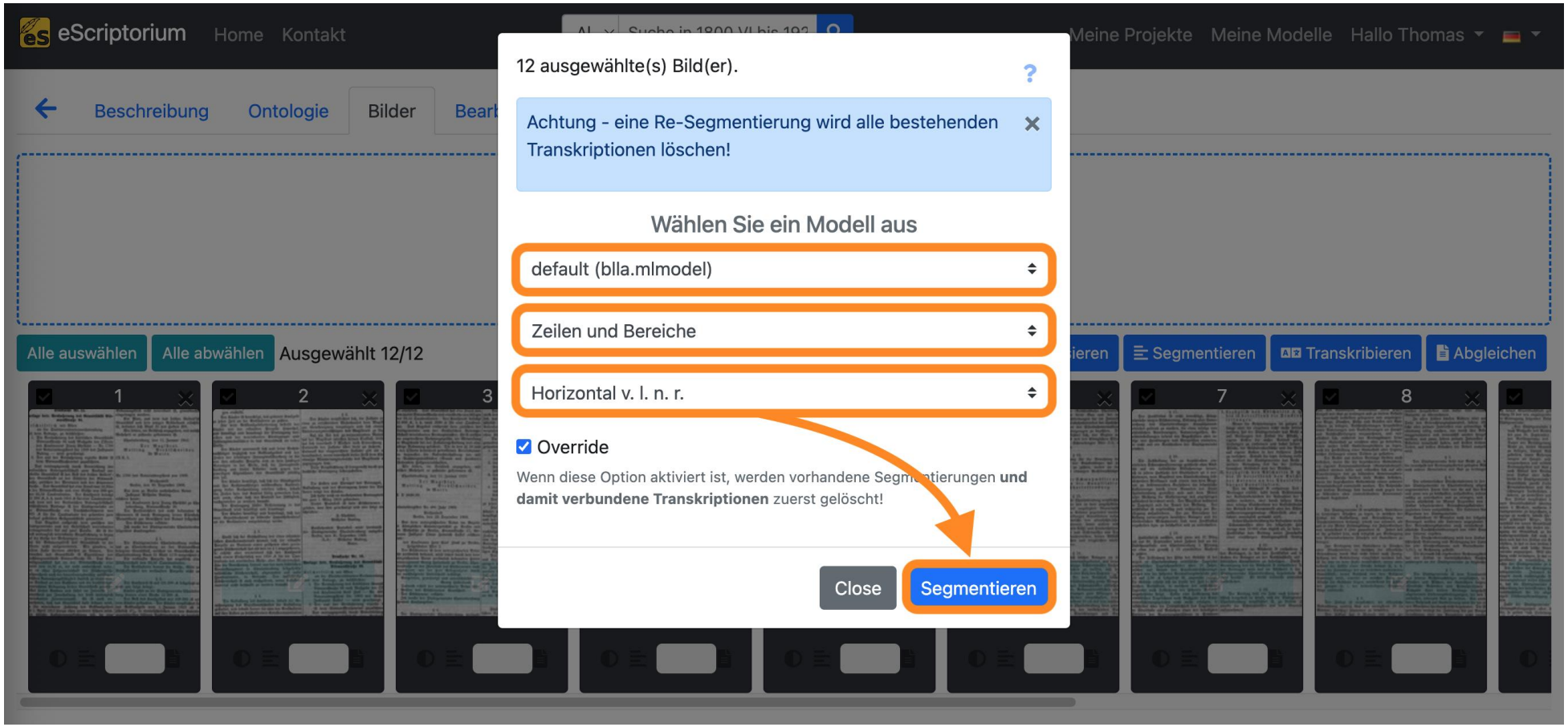


Theorie: Ergebnis der Segmentierung



Workflow-Schritt: Bild segmentieren

Workflow-Schritt: Bild segmentieren



Theorie: Ergebnis der Segmentierung

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

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/

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

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

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

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

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

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