Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Digitalisierung und Erschließung historischer Schulbücher mit Schwerpunkt religiöse Bildung – zwischen Masse und gezielten manuellen Eingriffen

Christian Reul

Zentrum für Philologie und Digitalität (ZPD)
Universität Würzburg

26.06.2024

Agenda

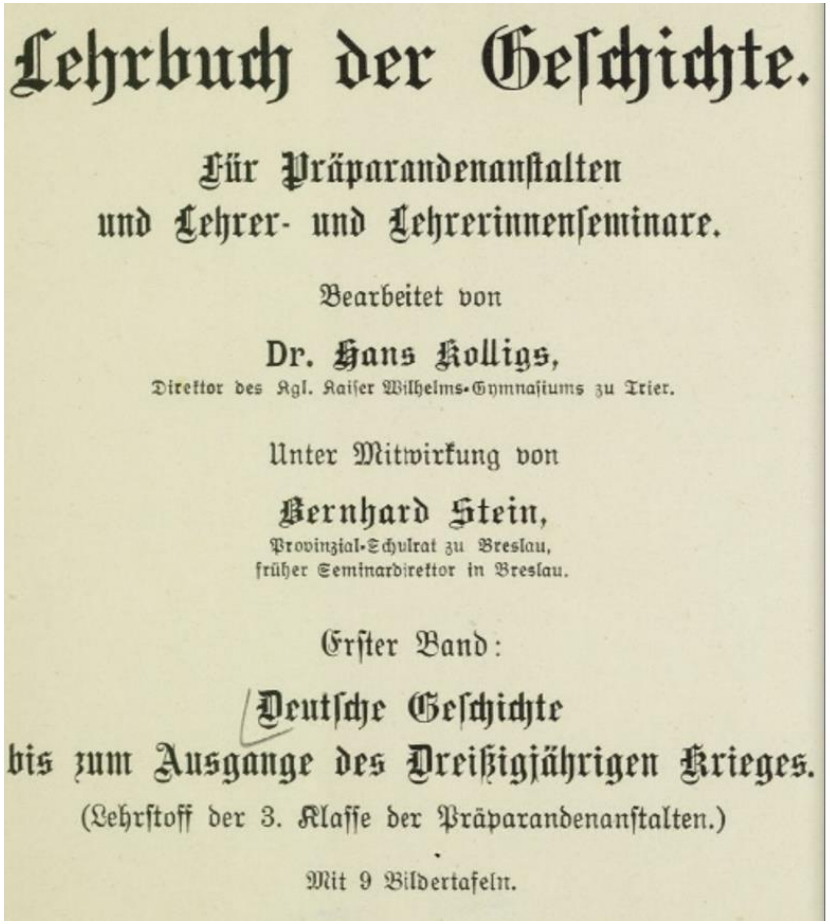
1. Allgemeine Projektinfos
2. Genereller (OCR) Workflow
3. Layoutanalyse
4. OCR-Qualitätssicherung und -Optimierung

Allgemeine Projektinfos

Digitalisierung und Erschließung historischer Schulbücher mit Schwerpunkt religiöse Bildung und Aufbau einer Wissensbasis für die bildungshistorische Forschung

- DFG-Projekt, Förderlinie „Digitalisierung und Erschließung“
- Kooperation mit Dr. Anke Hertling (Leibniz-Institut für Bildungsmedien | GEI)
- Laufzeit drei Jahre, Start 07/24
- Ziele:
  - Erfassung von deutschspr. Religionsschulbücher und schulischen (Erst-)Lesebüchern
    - Knapp 2.000 Bände mit fast 500.000 Seiten
    - Zeitraum 1618 bis 1870
  - Systematische Erfassung von Akteuren aus der historischen Schulbuchproduktion unter Einsatz von Named Entity Recognition (NER) und Named Entity Linking (NEL)

Beispiel NER-Ergebnis auf fehlerhafter OCR



Lehrbuch der Geſchichte.

Für Präparandenanſtalten
und Lehrer- und Lehrerinnenſeminare.

Bearbeitet von

Dr. Hans Kolligs,
Direktor des Kgl. Kaiſer Wilhelms-Gymnaſiums zu Trier.

Unter Mitwirkung von

Bernhard Stein,
Provinzial-Schulrat zu Breslau,
früher Seminardirektor in Breslau.

Erſter Band:

Deutſche Geſchichte
bis zum Ausgange des Dreißigjährigen Krieges.

(Lehrſtoff der 3. Klaſſe der Präparandenanſtalten.)

Mit 9 Bildertafeln.

Lehrbuch der Geſchichte MISC .

Für Präparandenanſtalten

und Lehrer- und Lehrerinnenſeminare.

Bearbeitet von

Dr. Hans Kolligs PER .

Direktor des Kal ORG . Kaiſer Wilhelms-Gymnaſiums ORG zu Ttier LOC .

Unter Mitwirkung von

Vernhard Stein PER .

Provinziol. Schulrat zu Breslau LOC ,

früher Eemiterdrettor in Greslas LOC .

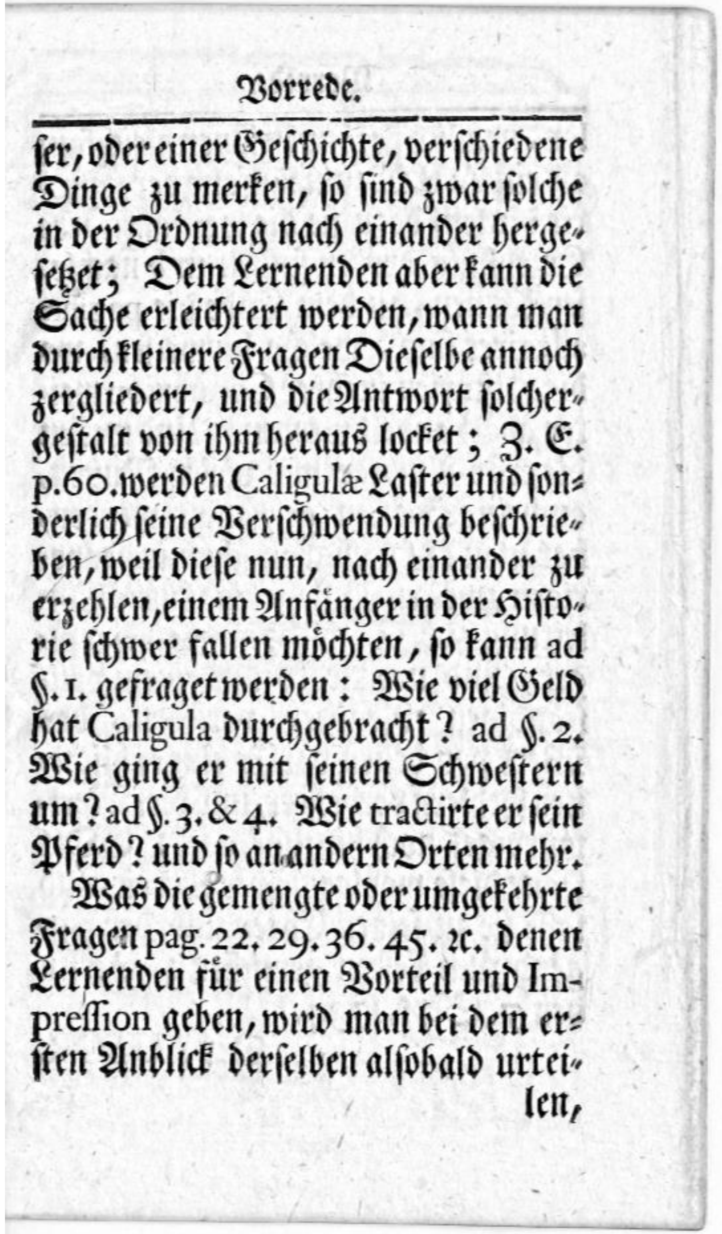
Erſter Band:

Deutſche Geſchichte

bis zum Ausgange des Dreißigjährigen Krieges MISC .

(Lehrſtoff der 3. Klaſſe der Präparandenanſtalten.)

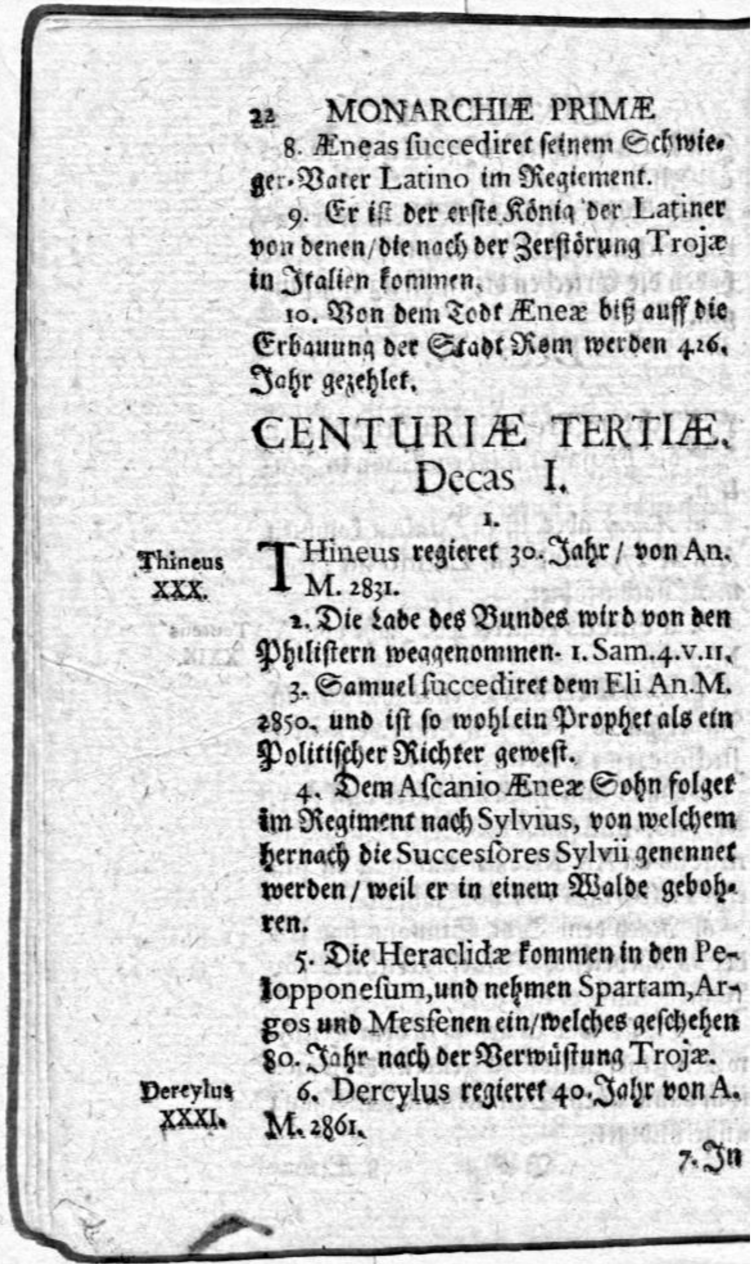
Mit 9 Bildertafeln.



Vorrede.

ser, oder einer Geschichte, verschiedene Dinge zu merken, so sind zwar solche in der Ordnung nach einander hergesetzet; Dem Lernenden aber kann die Sache erleichtert werden, wann man durch kleinere Fragen dieselbe annoch zergliedert, und die Antwort solchergestalt von ihm heraus locket; Z. E. p.60. werden Caligulæ Laster und sonderlich seine Verschwendung beschrieben, weil diese nun, nach einander zu erzehlen, einem Anfänger in der Historie schwer fallen möchten, so kann ad §. 1. gefraget werden: Wie viel Geld hat Caligula durchgebracht? ad §. 2. Wie ging er mit seinen Schwestern um? ad §. 3. & 4. Wie tractirte er sein Pferd? und so an andern Orten mehr.

Was die gemengte oder umgekehrte Fragen pag. 22. 29. 36. 45. 2c. denen Lernenden für einen Vorteil und Impression geben, wird man bei dem ersten Anblick derselben alsobald urteilen,



22 MONARCHIÆ PRIMÆ

8. Æneas succediret seinem Schwieger-Vater Latino im Regiement.

9. Er ist der erste König der Latiner von denen / die nach der Zerstörung Trojæ in Italien kommen,

10. Von dem Todt Æneæ biß auff die Erbauung der Stadt Rom werden 426. Jahr gezehlet.

CENTURIÆ TERTIÆ.

Decas I.

1.

Thineus XXX. Thineus regieret 30. Jahr / von An. M. 2831.

2. Die Lade des Bundes wird von den Philistern weggenommen. 1. Sam. 4. v. 11.

3. Samuel succediret dem Eli An. M. 2850. und ist so wohl ein Prophet als ein Politischer Richter gewest.

4. Dem Ascanio Æneæ Sohn folget im Regiment nach Sylvius, von welchem hernach die Successores Sylvii genennet werden / weil er in einem Walde gebohren.

5. Die Heraclidæ kommen in den Peloponnesum, und nehmen Spartam, Argos und Messenen ein / welches geschehen 80. Jahr nach der Verwüstung Trojæ.

Dercylus XXXI. 6. Dercylus regieret 40. Jahr von A. M. 2861.

7. In

Des XVII. Seculi — Der OESTERREICHISCHEN KAISER — Tab.

DEUTSCHE KAISER.

DES DEUTSCHEN REICHS STÄNDE.

Anmerkungen.

Genereller (OCR) Workflow

- Bearbeitung mittels OCR4all und LAREX
- Binarisierung wohl mit ocrd-sbb-binarize

Gibt‘s was neueres/besseres?

- Layoutanalyse → nächste Folie
- Erkennung voraussichtlich mit Calamari, vorab allerdings noch Tests; Neue gemischte Modelle (hoffentlich) bis dahin fertig

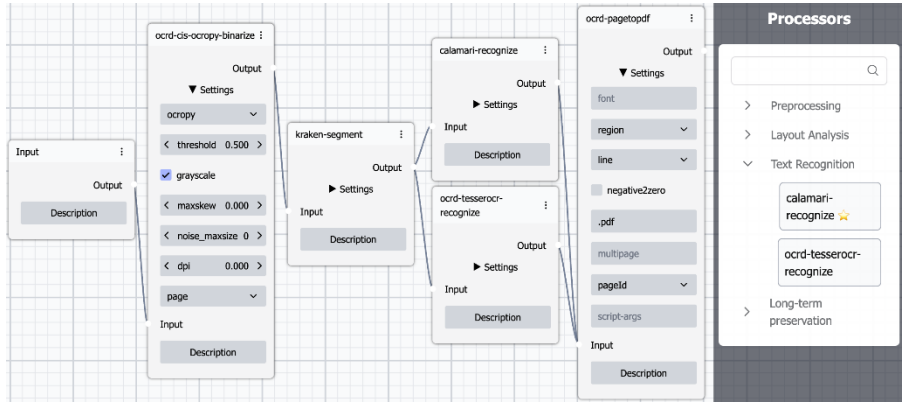
Tester gesucht!

- OCR-Qualitätssicherung und -Optimierung → überübernächste Folie

- NER mit (angepassten) Standardlösungen
- Unterstütztes (GND-API + eigene GUI) NEL

Layoutanalyse

- Ursprünglicher Plan: OCR-D Workflow Suche
  - Erstellung via NodeFlow
  - Ergbnisabgleich via LAREX
  - Gezielter Einsatz via OCR4all-Tagging-Funktionalität



- Aber: lieber direkt voll auf trainierbare Methoden setzen
  - Besser und nachhaltiger
  - Paralleles Projekt zur Layoutanalyse (mit Stabi Berlin und SLUB Dresden) → Synergieeffekte

→ Fokus auf Kraken und (evtl. später) Eynollah

→ gezielt Trainingsdaten erstellen, Active Learning

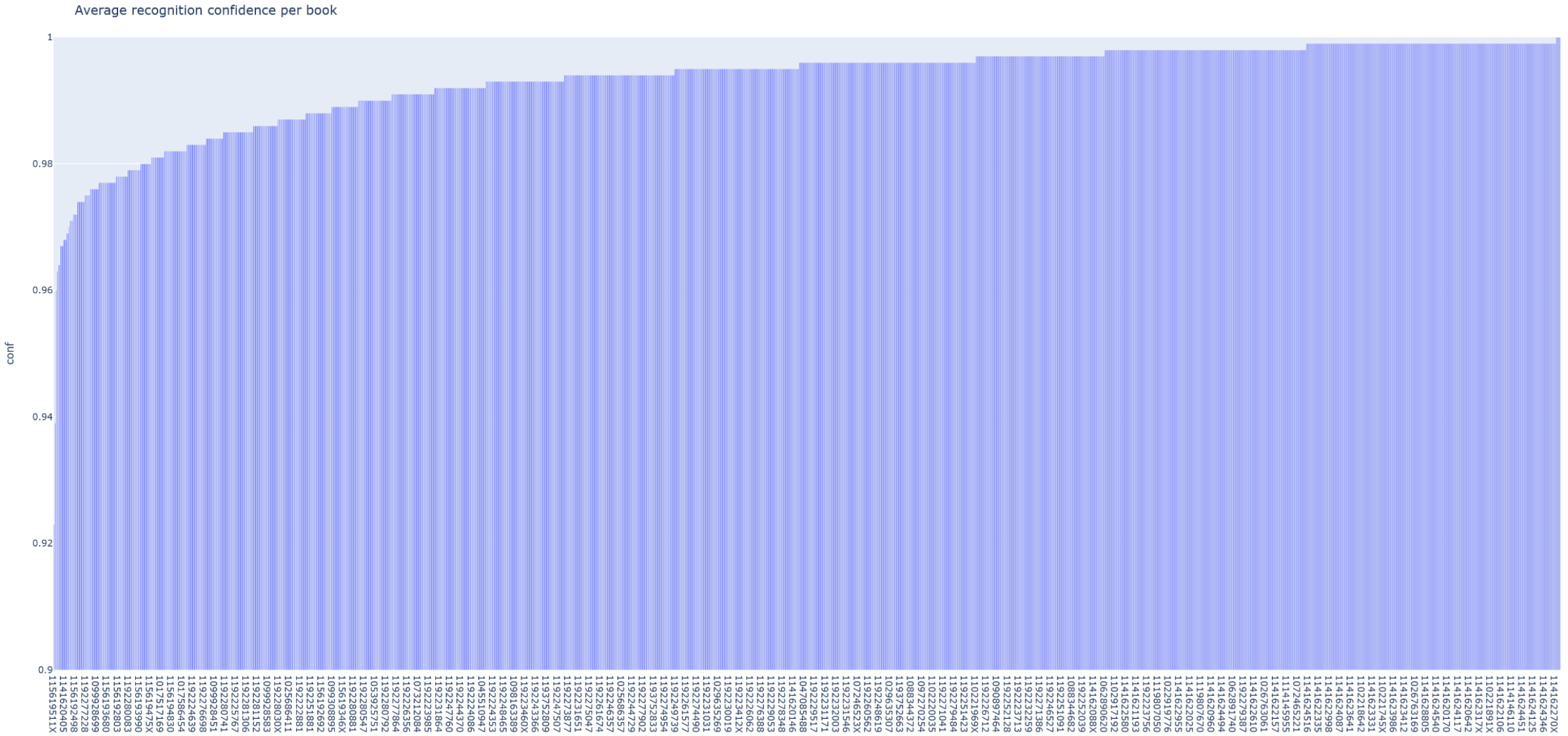
Gerne mitmachen! (mit eigenen Daten)

OCR-Qualitätssicherung und -Optimierung

- Eingehende manuelle Prüfung aller Werke/Seiten nicht gangbar
- Stattdessen gezielte Prüfung, basierend auf Qualitätsabschätzung
  - Auf Werk-/Seitenebene → jeweils sortiert nach (vermuteter) „Qualität“
  - Flexibel und experimentell!
- Art des Eingriffs dann je nachdem
  - Anpassung von Parametern, Anwendung eines anderen Modells, …
  - Erstellung von Trainingsdaten und gezieltes Nachtraining (auf Werk-/Clusterebene)
  - …?
- Frage: Wie Qualität abschätzen?
  - Bereits gute Erfahrungen mit OCR-Konfidenz, allerdings auch klare Schwächen
  - Nur pragmatische (und idealerweise direkt einsetzbare) Lösungen sinnvoll

Erfahrungen? Ideen?

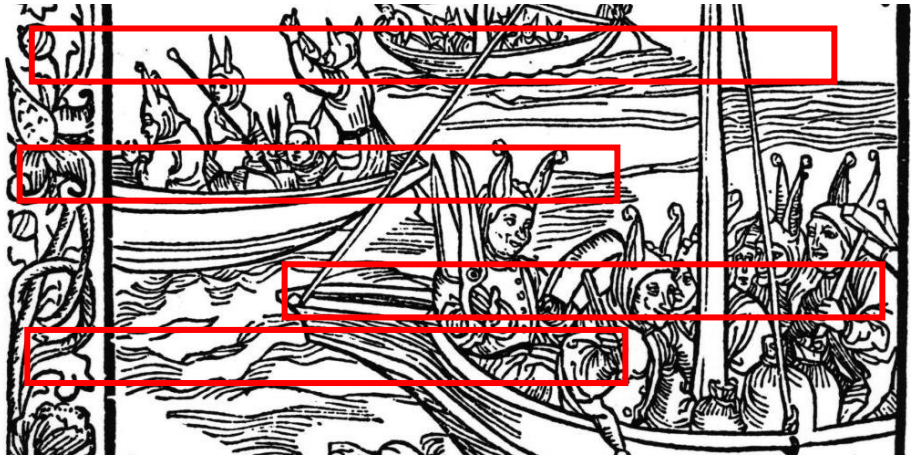
Anwendungsfall „Fotidis‘ Schundhefte“



Stärken/Schwächen OCR-Konfidenz

Text Text Text Text Text

teilsegmentierte Zeile, Zeilen in Bild
→ OCR-Konfidenz klarer Indikator



Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam

Übersehene Spaltentrennung
→ OCR-Konfidenz ggf. top,
Ergebnis jedoch Quatsch