

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





Agenda

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



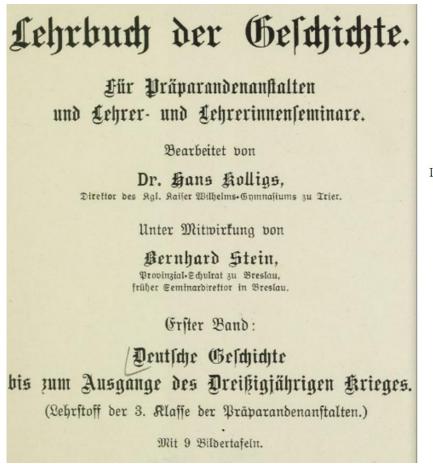
Allgemeine Projektinfos

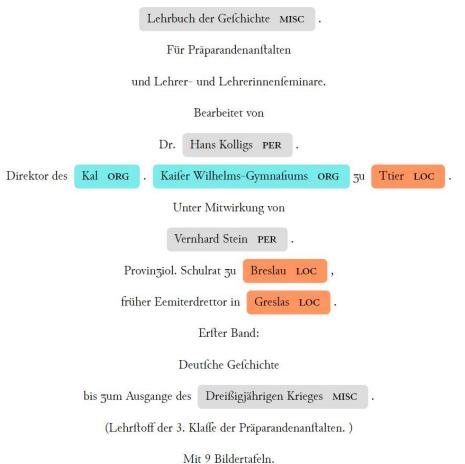
Digitalisierung und Erschließung historischer Schulbücher mit Schwerpunkt religiöse Bildung und Aufbau einer Wissensbasis für die bildungshistorische Forschung

- 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







Was die gemengte oder umgefehrte Fragen pag. 22, 29.36. 45. 2c. denen Lernenden für einen Borteil und Impression geben, wird man bei dem er= sten Unblick derselben alsobald urteiMONARCHIÆ PRIMÆ

8. Aneas fuccediret feinem Ochwie. ger. Bater Latino im Regiement.

9. Er iff ber erfte Ronig ber Latiner bon benen/bie nach ber Berfforung Trojæ in Stalien fommen,

10. Bon bem Tedt Enem bif auff bie Erbauung ber Ctabt Rom werben 426. Jahr gezehlet.

CENTURIÆ TERTIÆ. Decas I.

Thineus XXX.

T Hineus regieret 30. Jahr / von An.

2. Die Labe bes Bundes wird von ben Philiftern weggenommen. I. Sam.4.v.II.

3. Samuel fuccediret bem Eli An.M. 2850. und ift fo mohlein Prophet als ein Politifcher Richter geweft.

4. Dem Afcanio Enea Cohn folget im Regiment nach Sylvius, von welchem bernach die Successores Sylvii genennet werden / weil er in einem Walde geboh.

5. Die Heraclidæ fommen in ben Pe-Topponesum, und nehmen Spartam, Argos und Mesfenen ein/welches gefchehen 80. Jahr nach der Bermuftung Trojæ.

6. Dercylus regieret 40. Johr von A.

ifchen Kbaige Christiano IV. ber an Schie ber a. 26. ben Butter geschlagen verch.

if inbeffen glibdlich meber bie Tar-

afig s. 94, folgt, ber facholift in Straig in Bolen wird. In ber Erneftmifden Piete fürft in Meimar fol

potente and parties (1) potente and parties (1) potente arterial potente (1) potente and potente (1) p

Menthe tille a 15.

5. FERDINANDUS II, med bett

N. MEDESCHINENDUS III selbert

N. MEDESCHINENDUS

N. MEDESCHINENDUS

N. MEDESCHINENDUS

N. MEDESCHINENDUS

N. MEDESCHINENDUS

N. MEDESCHI

Ehotwarde ethalten ber n. 32. der George Laderican fele 1876-Flenburg. Die herzoge werben a. 12. in der Alde erflatzung Wolfenfach läßt fich a. 12. albie bulbieren. 2 lein der König im Schrecken verlieseriffen, 31. und sie nehmen 2. 20. den Pragischen Areden an. Adolphus Frideri yn Schressingeriffen, Dader Sohn Christians Lodoscos mit Ausschließung der Beilder folge. Er wird in Fra erich 2. 63. Latholisch, und Größe 3. in Hollund ohne Erben.

) Collifein. In Gottorp finder Frederium III. a. 35. eine Gefandichaft in Perfen, und erlangt n. 56. in Schlein Sworenziginit, furdit a. 49. Sein Sohn Christianus Albertus fliebi a. 73, und ca. für Danemark, wird aber a. 50. in die Bartheren. Der Ra finniet, nab firibt a. 94. Defen Caba Fridericus IV. priult abermal mit Danemarf,erlangrober a. 1700. Den Trave

Bill Conference Luiser, and program deligher Arriver.

Mit Conference Luiser, and December Homes im Weithelischen Frieden an Scharfers. Magnetium was Halberfladt aber an Translature.

1. Arti Conference and Scharfers. And A. Hinfolder liste description of Like Conference and Conference and

echet, mid a st. barg und Pfell Berbarg wegen ibrer Bemohlnuer gagregen, und jich endlich a. 24. vergleichen. rung von Mache () Minden femmt im manfteriden Frieden au Brandenburg.

around ben Kriefl des Lives of Oceantries a siche street seem and the Chinese of Oceantries as siche seems of Oceantries as oceantries of Oceantries as oceantries of Oceantries as oceantries of Oceantries as oceantries oceantri refined Constitue has December and principles in the Namipies, or imposed time Official and Nami A. 2, a 19 ed 20 ed. 2, a 19 ed. 2, genfeldinn die Kaustrafen, wit herrat het feine Britagin en den dur die Orleans, med fürde ". 30., 4. 8 5. wormt die Registung an Philippen Wilhelmum aus der Entheligiere Britannif fein Tin

The Company of the Co FERDINANDES III. jum b) Geffert. Maartinispositel immer untersent entante erroute en contente. Annum in please de please de la contente del la contente de la contente del la contente de la contente del

of Merydine, machanis o protest.

of Merydine, machanis o protest.

of April 1985.

of Merydine, machanis of April 1985.

of Merydine, machanism of Apr nd mit Cadijen. Weungel fallt in Difficiburg, Cionen, Dies und Ondumar. Dach erlangt bie Ctatibaltericaft in Welfrieftant. Die 2B.

ind in Californ. We make the are Dilladoring Science 200 and 2

A. 1826-Object. Extribited on Ordering a commonwer position protect extract (chairmen and positions before he was finite present recognizing from the protect protect of the protect protect of the protect protect of the protect pro

Northern Wester. Die Erfert in der Geschen und der Geschen der Ges

A BURGUNDISCHE RREIS. A. 9. machen Dit.7. Provenzen mit Spanien einen zujährigen Stillkant ritus fürbt a. 24. nab ber Bruder Feidericus Henricus a. 47. Der Caba Wilhelmus II. foldt in der Sta inft de merben, 45, der gewinger von Sparien med bem Anfer pres erfahrt. Er gar bier dreutfelinften School, und unfera, 50, de bien mich dem Lobe der Pries Wilhelm geloberen weir. Die Cattholierficheft nieb gear a. 52, annulier, inder 2, 72, and Ermordung der Gefebber Wieren, Wilhelm micher anglertenen.

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

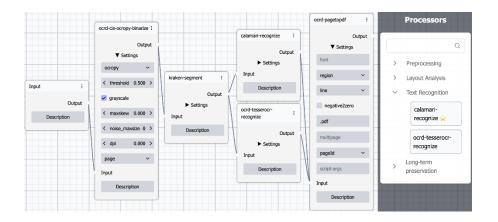
Tester gesucht!

- 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 <u>Projekt zur Layoutanalyse</u> (mit Stabi Berlin und SLUB Dresden) → Synergieeffekte
 - → Fokus auf Kraken und (evtl. später) Eynollah
 - → gezielt Trainingsdaten erstellen, Active Learning

Gerne mitmachen! (mit eigenen Daten)



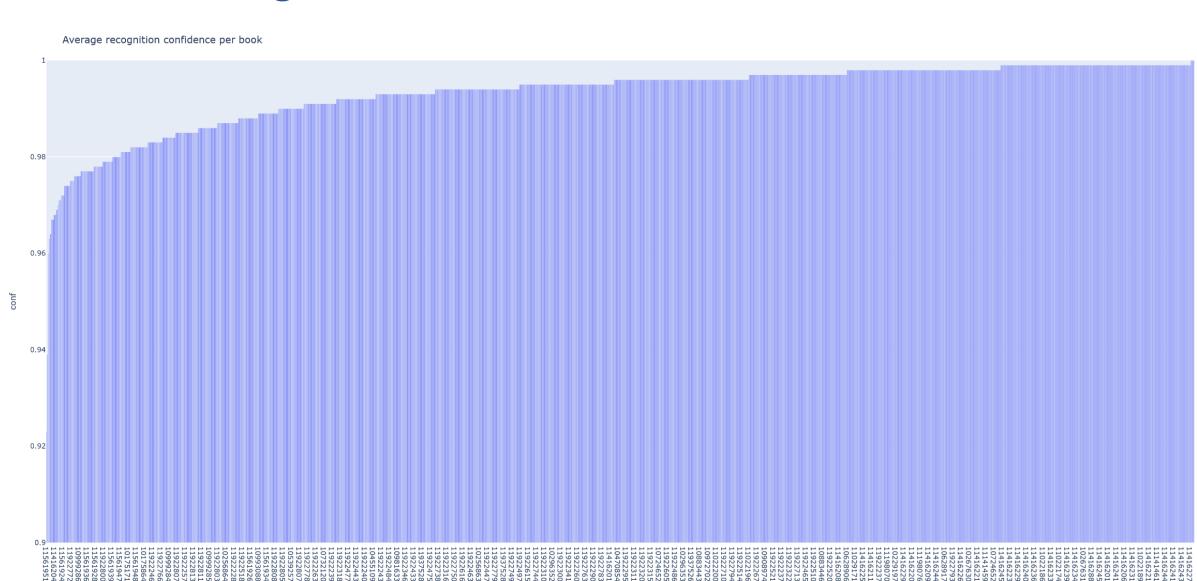
OCR-Qualitätssicherung und -Optimierung

- 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?



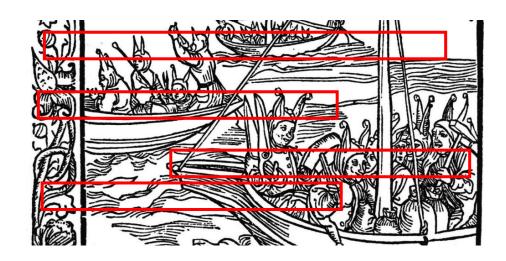
Anwendungsfall "Fotidis' Schundhefte"



Stärken/Schwächen OCR-Konfidenz

teilsegmentierte Zeile, Zeilen in Bild

→ OCR-Konfidenz klarer Indikator



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

Übersehene Spaltentrennung

→ OCR-Konfidenz ggf. top, Ergebnis jedoch Quatsch

