Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

LAREX - Ein Werkzeug zur Layout-Analyse und Segmentierung von frühen Buchdrucken

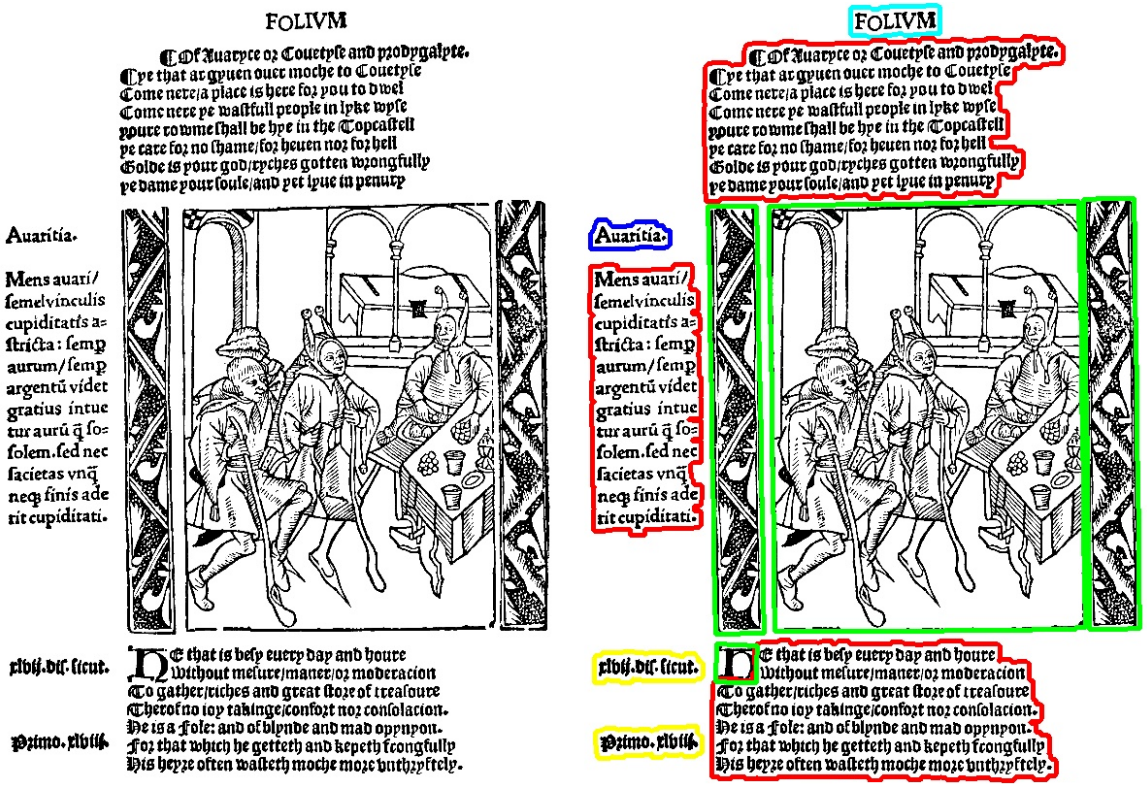
Christian Reul

Lehrstuhl für Künstliche Intelligenz und Angewandte Informatik
Universität Würzburg

17.03.2017

Motivation und Ansatz

- Segmentierung vor OCR meist unerlässlich.
- Text/Bild-Trennung oft nicht ausreichend.
- Häufig zusätzliche semantische Auszeichnung erwünscht.
- Semi-automatisch.
- Intuitiv, anpassbar und nachvollziehbar.
- Open Source.
- Layout Analysis and Region EXtraction.



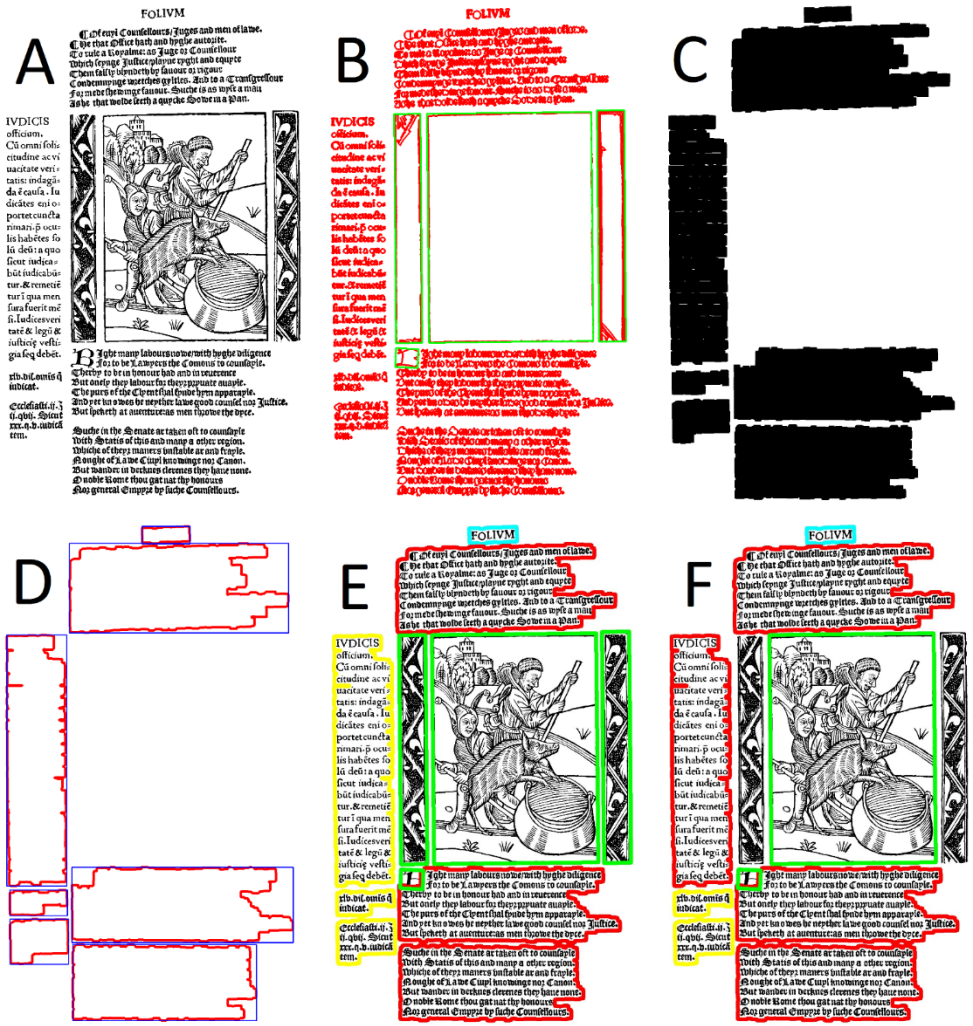
Input

Beispiel-Output

Workflow

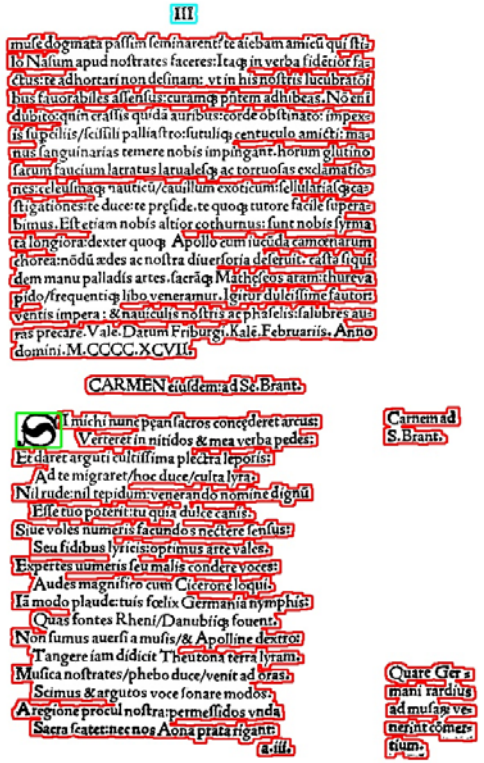
Input: Scan einer Seite.
Output: Klassifizierte Segmente als PageXML.

1. Pre-processing (A).
2. Bilddetektion (A → B).
    1. Optionale Dilatation des Vordergrundes.
    2. Bild/Nicht-Bild-Klassifikation.
    3. Entfernen der Bilder.
3. Grobe Textklassifikation (C → D, E).
    1. Dilatation des Vordergrundes.
    2. Regelbasierte Klassifikation.
4. Manuelle Korrektur und feine Textklassifikation (E → F).
5. Konvertierung in Ausgabeformat.

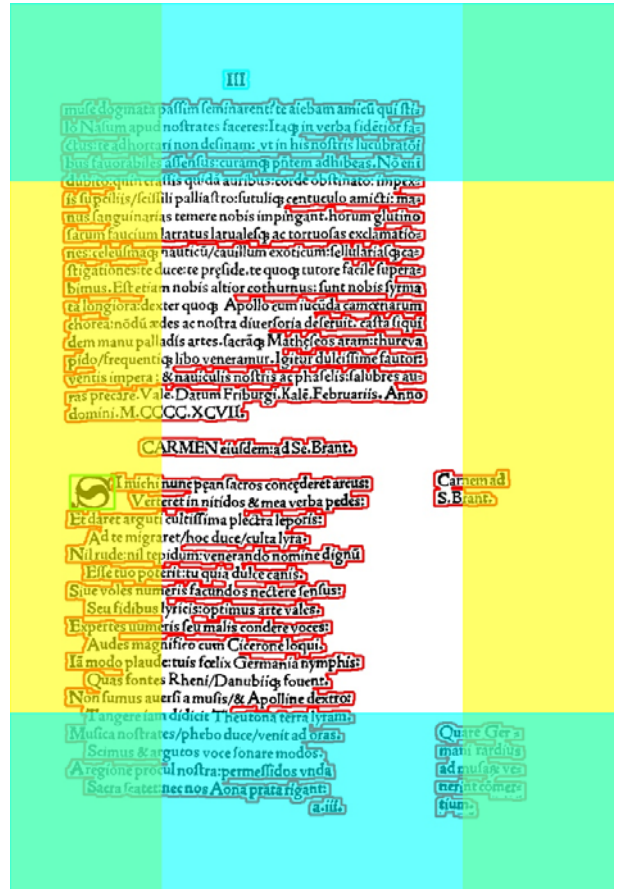


Globale Optimierung - Überblick

- Einzelne Bücher meist homogen.
  - Abstände.
  - Positionen (z. B. von Marginalien).
  - …
- Idee: Finden einer passenden, individuellen Layout-Maske.
- Manuelle Optimierung durch
  - Anpassung der Dilatations-Parameter.
  - Hinzufügen und Verschieben von Regionen-Typen und deren Positionen.
  - Anpassung der Regionen-Mindestgröße.
  - …



Segmentierungsergebnis mit Default Setup



Default-Positionen für Seitenzahl (cyan) und Marginalien (gelb).

Globale Optimierung - Dilatation

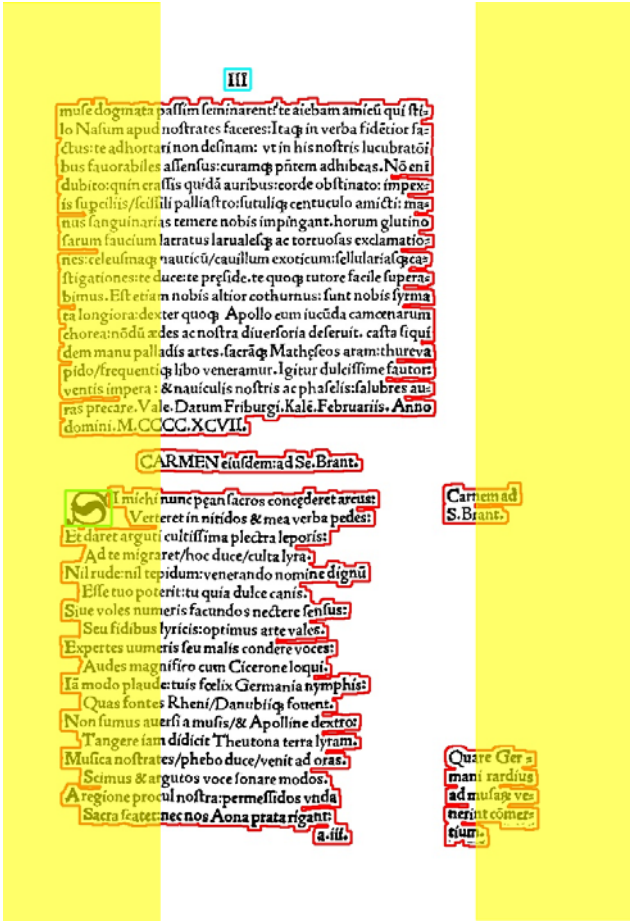
- Erweiterung von Vordergrundpixeln um X/Y Pixel in x/y-Richtung.
- Gut passender Wert essentiell:
  - Zu niedrig: Trennung zusammen gehöriger Zeichen/Wörter/Zeilen.
  - Zu hoch: Verschmelzung nicht zusammen gehöriger Blöcke.
- Optimierung im Normalfall sehr einfach.

Default Parameter

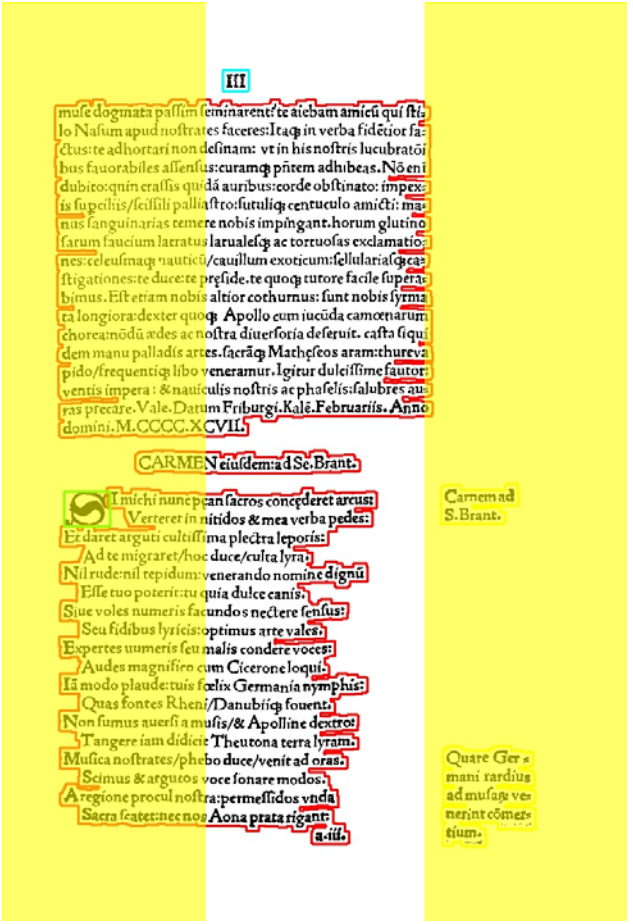
Optimierte Parameter

Globale Optimierung - Positionen

- Zuweisung nur bei vollständiger Überdeckung eines Blocks.
- Beliebige Anzahl von Positionen.
- Komfortables Hinzufügen, Löschen und Verschieben möglich.

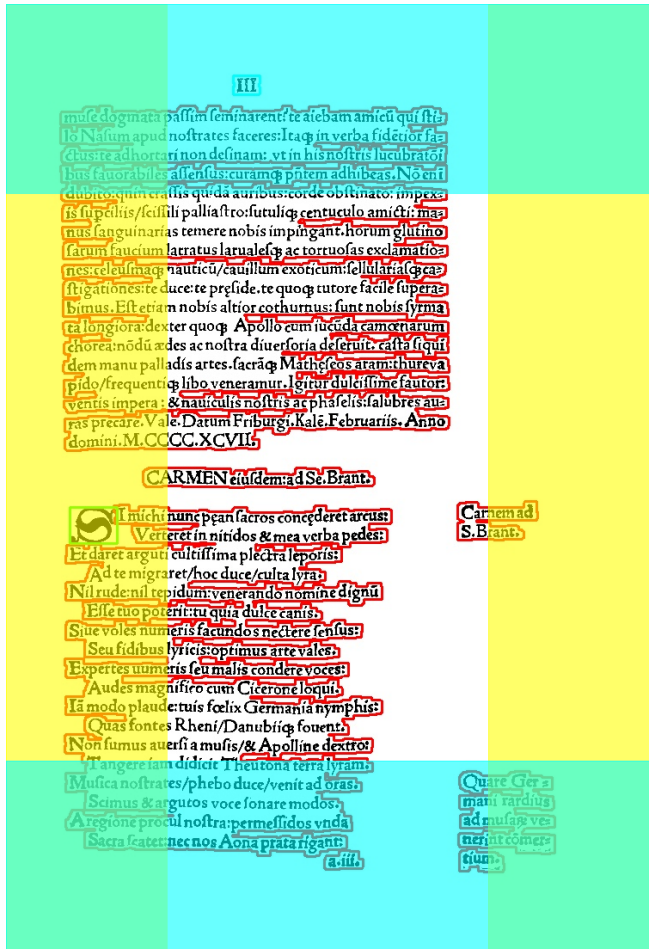


Default Positionen

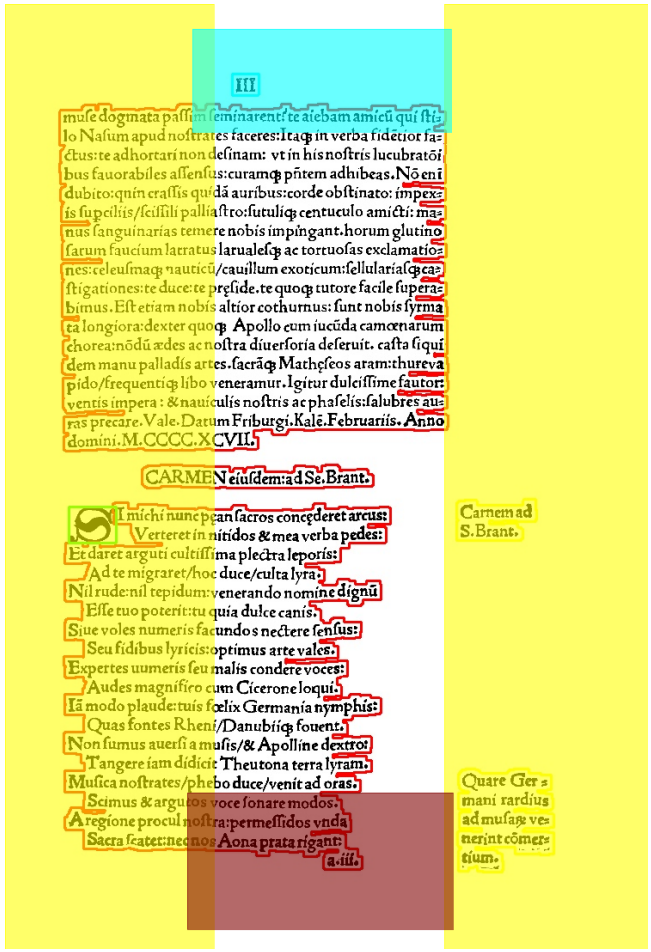


Optimierte Positionen

Globale Optimierung - Ergebnis



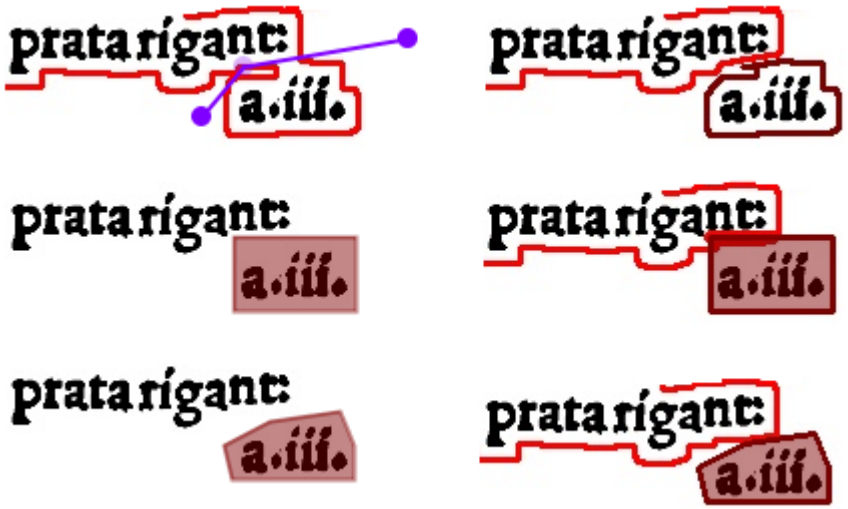
Default Setup und Ergebnis



Optimiertes Setup und Ergebnis

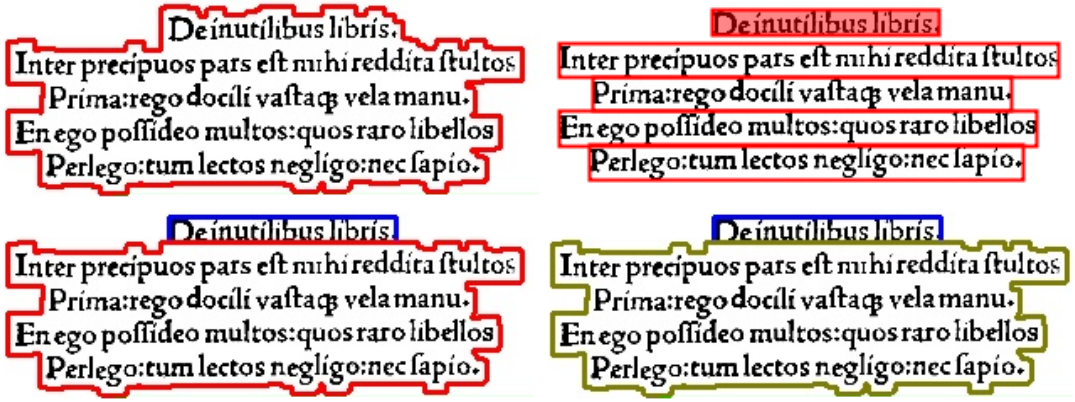
Lokale, manuelle Korrekturen

- Löschen von Regionen.
- Ändern des Typs.
- Blocktrennung durch Linien.
- Manuelle Auszeichnung von Regionen durch Rechtecke oder Polygone.

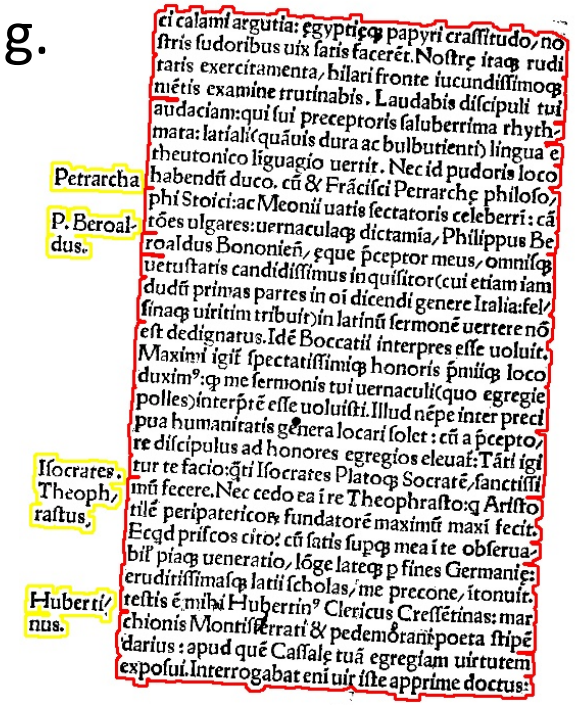


Feinsegmentierung auf Zeilenebene

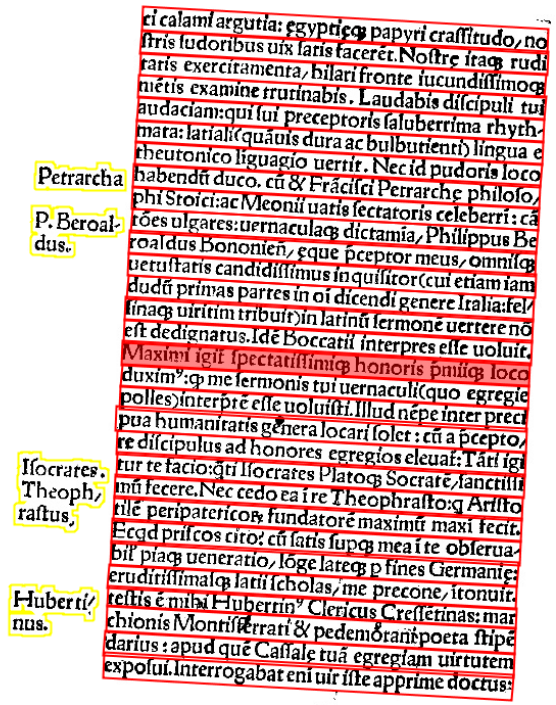
- Nutzung der Tesseract Zeilensegmentierung.
- Anschließend Blocktrennung anhand der Zeilen möglich.
- Zeitintensiv (Sammeln!).



Blocksegmentierung (ol), Zeilensegmentierung (or), markierte Überschrift (ul), Endergebnis (ur).



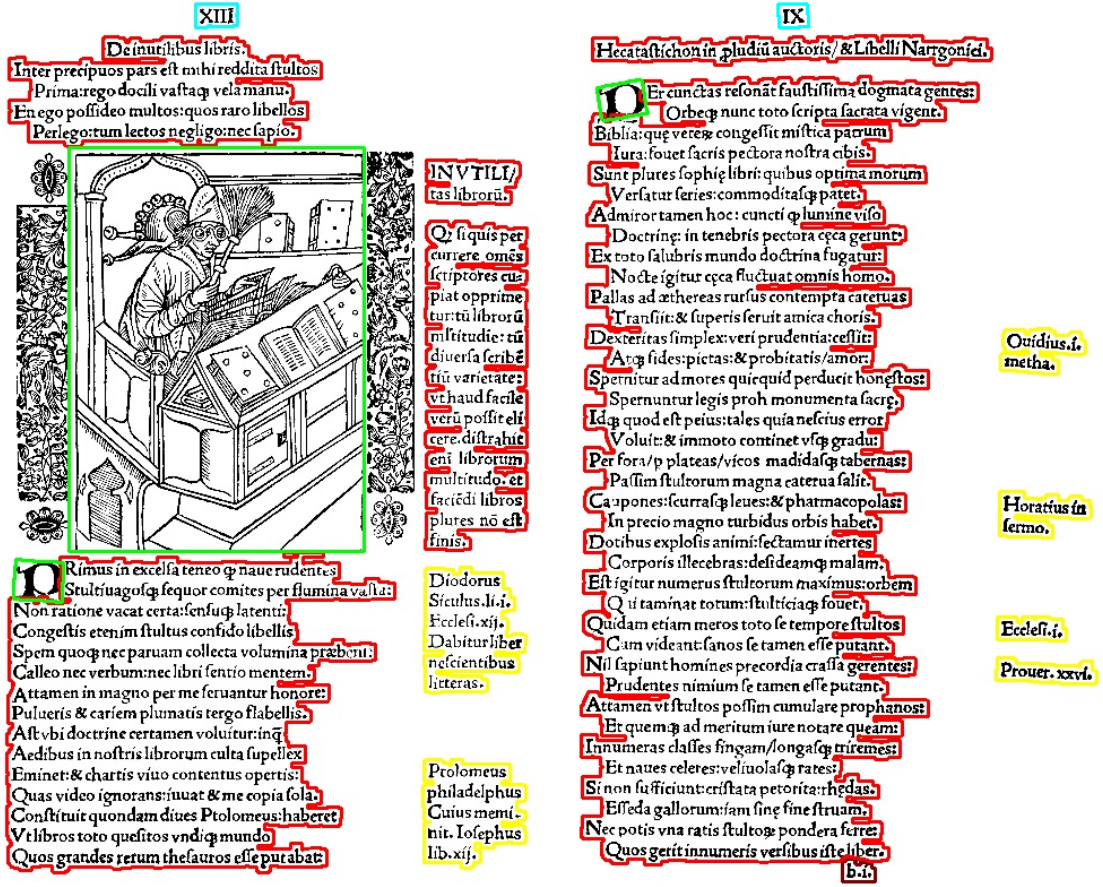
Ergebnis Blocksegmentierung



Ergebnis Zeilensegmentierung

Evaluation I – Barclay I

- Werk: Barclay‘s Narrenschiff.
  - Nachdruck von 1509.
  - 570 Seiten.
- Zeitaufwand:
  - LAREX: Komplettes Werk in 2h 18min.
  - Aletheia: In gleicher Zeit nur 160 Seiten.



Erwartete Segmentierung zweier Beispielseiten

Evaluation I – Barclay II

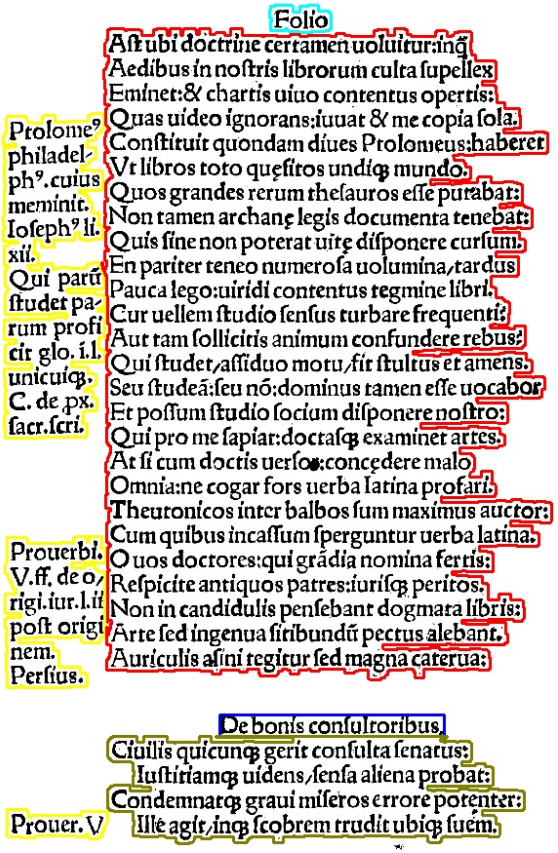
- Erfassung der benötigten Änderungen auf den ersten 200 Seiten.
- Sehr simple Änderungen überwiegen.
- Spezielle Nutzeranforderungen können Aufwand erheblich steigern.

Art der Änderung	#	Aufwand
Irrelevante Region entfernen	161	Rechtsklick
Irrelevantes Bild entfernen	90	Rechtsklick
Typkorrektur: Marginalie → Bildbeschreibung	65	Doppelklick und Typ
Abtrennen der Bogensignatur	31	Linie zeichnen
Typkorrektur: Fließtext → Marginalie	17	Doppelklick und Typ
Verbinden zweier Textblöcke	10	Rechteck und Typ
Manuelles Auszeichnen eines Bildes	2	Rechteck und Typ
Manuelles Auszeichnen der Seitenzahl	2	Rechteck und Typ

Evaluation II – Reale Anwendung auf *Narrenschiffe*

- Im Rahmen des *Narragonien digital*-Projekts.
- Erzielte Ergebnisse:

	Barclay	GW5064	GW5066
Anzahl Seiten	570	351	234
DPI	400	400	96
Grobsegmentierung	2h 18min	2h 20min	2h 00min
Feinsegmentierung	2h 40min	2h 15min	1h 45min
Zeichengenauigkeit auf Testset	98.40%	98.13%	98.80%



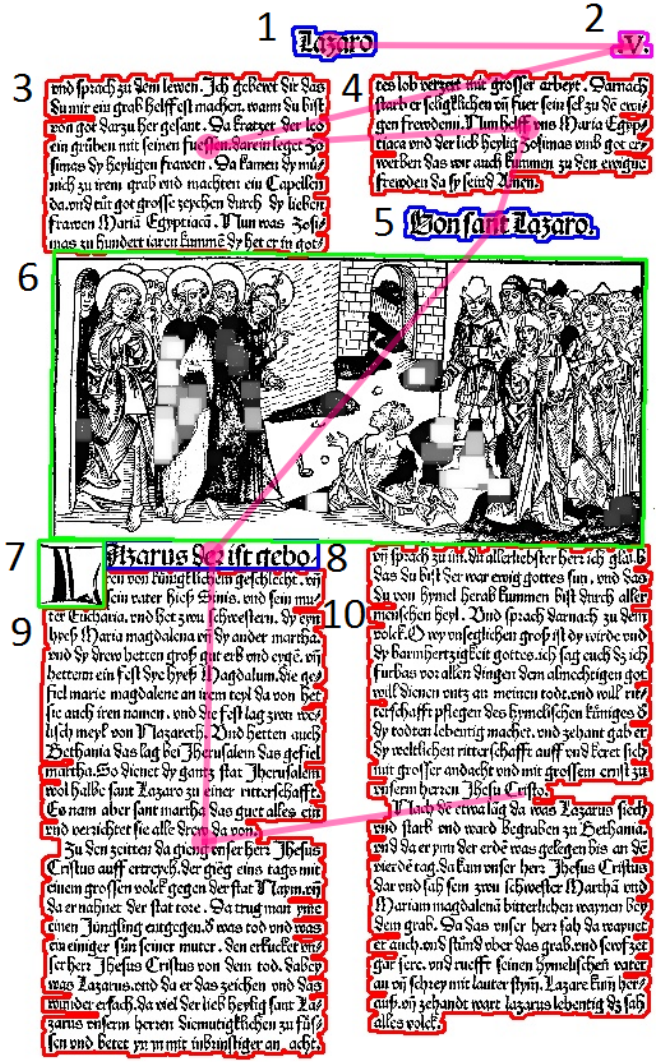
Beispielseite GW5064

Beispielseite GW5066

Evaluation III – Der Heiligen Leben

- Test einer vollautomatischen Segmentierung.
- Werk: I.t.f. 954 – Der Heiligen Leben (1488).
- Vorherige Anpassung auf Codeebene nötig!
- Vergleich des Zeitaufwands der Segmentierung und des resultierenden OCR-Ergebnisses.
- Ergebnisse:

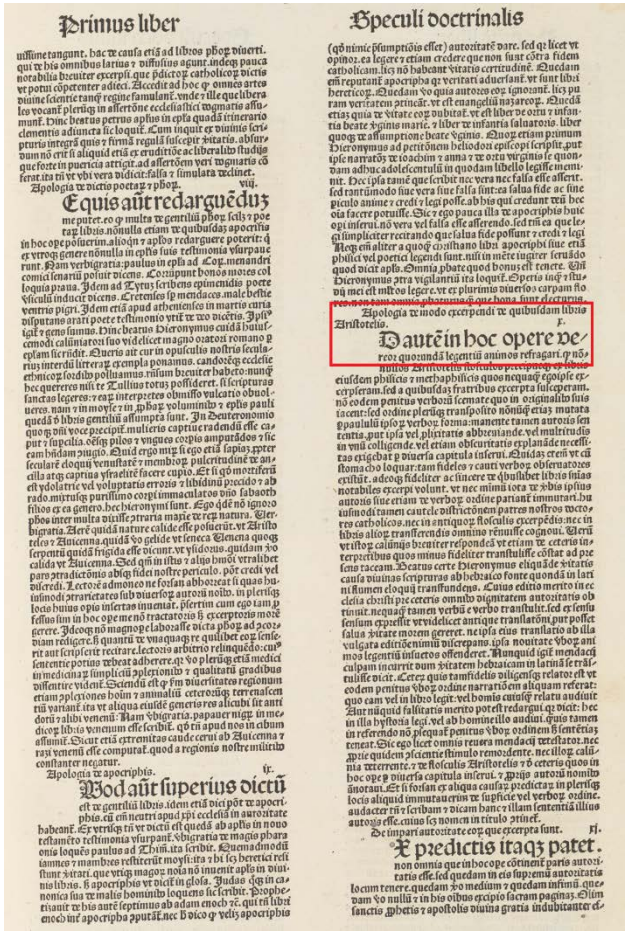
	Aletheia	LAREX
Manueller Zeitaufwand	ca. 100h	ca. 1h
Zeichengenauigkeit (95% KI)	97,57 (20) %	97,35 (28) %
Wortgenauigkeit (95% KI)	92,19 (51) %	91,84 (62) %



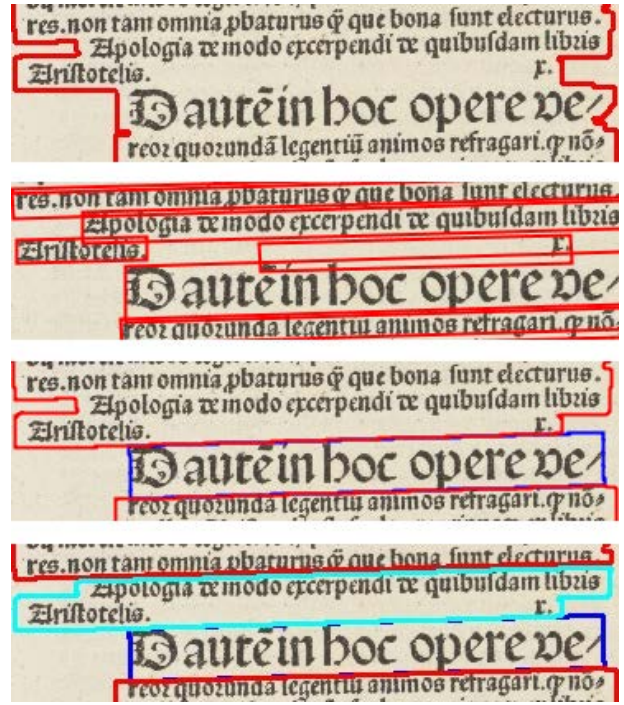
Korrekte Segmentierung einer Beispielseite

Aktuell: Vinzenz von Beauvais‘ Specula

- „Enzyklopädien“ des Spät-mittelalters (Druck 1481-1486).
- Vier umfassende Werke (knapp 3300 Seiten).
- Homogenes, zweispaltiges Layout…
- … leider mit ein paar Extras.

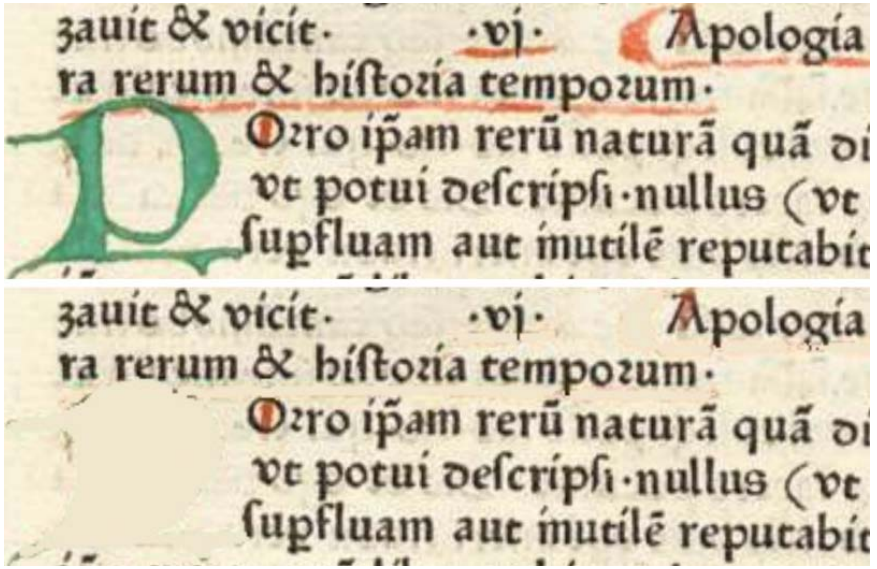


Beispielscan (beschnitten)


Schritte einer möglichen Segmentierung

Diskussion und Ausblick

- Intuitives Tool zur semi-automatischen Segmentierung von frühen Buchdrucken.
- Bereits sehr gute Ergebnisse in realen Anwendungsszenarien.
- Trotzdem noch viele Baustellen:
  - (Web-)Oberfläche.
  - Ausbau des Regelsystems.
  - Adaptation an neuere Drucke.
    - (Optionale,) robustere Bilderkennung.
    - Text in „Bild“.
    - …
  - Integration Vorverarbeitung.
  - Integration Initialenentfernung.
  - …



Beispiel Initialenentfernung

Vielen Dank für Ihre Aufmerksamkeit!

- Reul, C., Springmann, U., Puppe, F.: LAREX – A semi-automatic open-source Tool for Layout Analysis and Region Extraction on Early Printed Books.
  - Eingereicht bei DATeCH 2017. Verfügbar auf arXiv.org.
  - Tool Homepage (im Aufbau).
- Reul, C., Dittrich, M., Gruner, M.: Case Study of a highly automated Layout Analysis and OCR of an incunabulum: ‚Der Heiligen Leben' (1488).
  - Eingereicht bei DATeCH 2017. Verfügbar auf arXiv.org.
  - Original Scans ‚Der Heiligen Leben' (Virtuelle Bibliothek Uni Würzburg).
  - OCR-Ergebnis der 41 Testseiten.