Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Kallimachos

<philtag n="13"/>

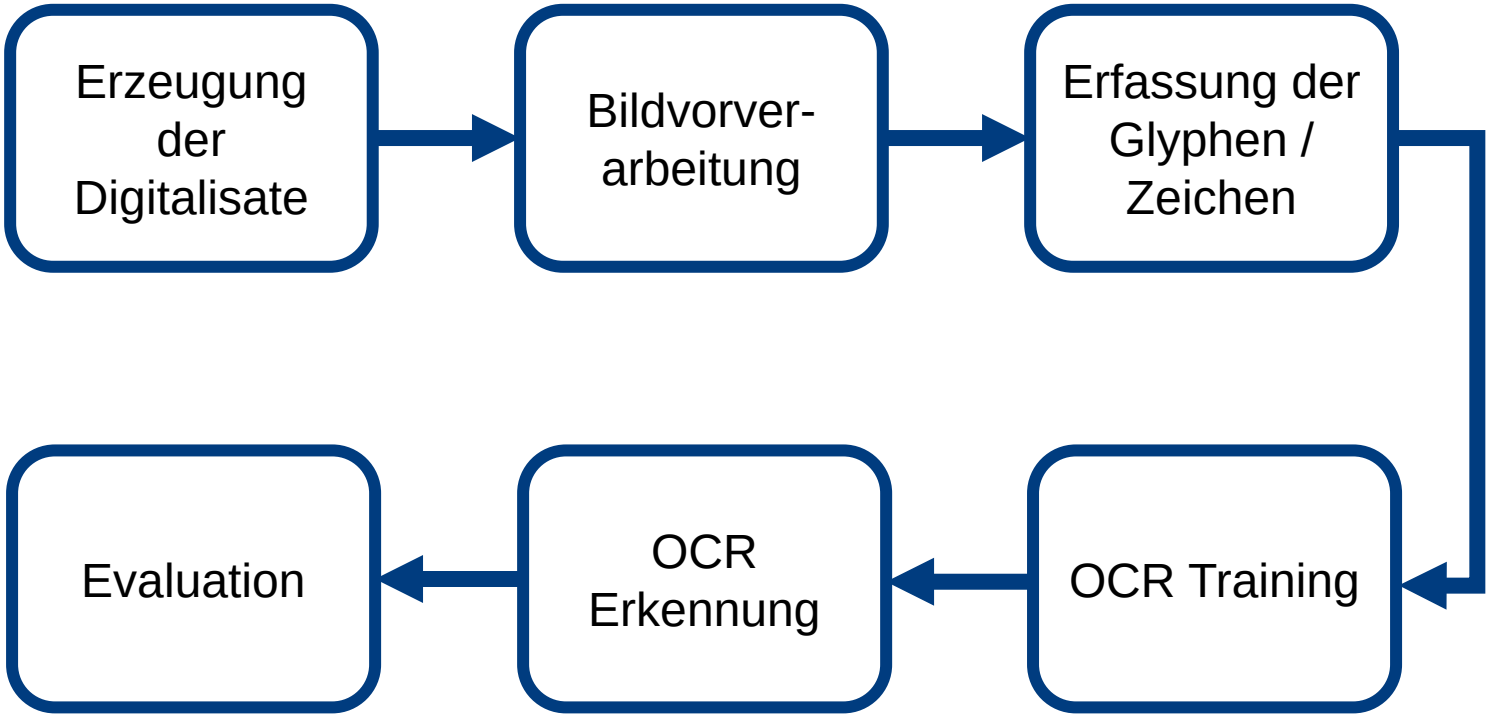
OCR Workshop

Marco Dittrich, Felix Kirchner

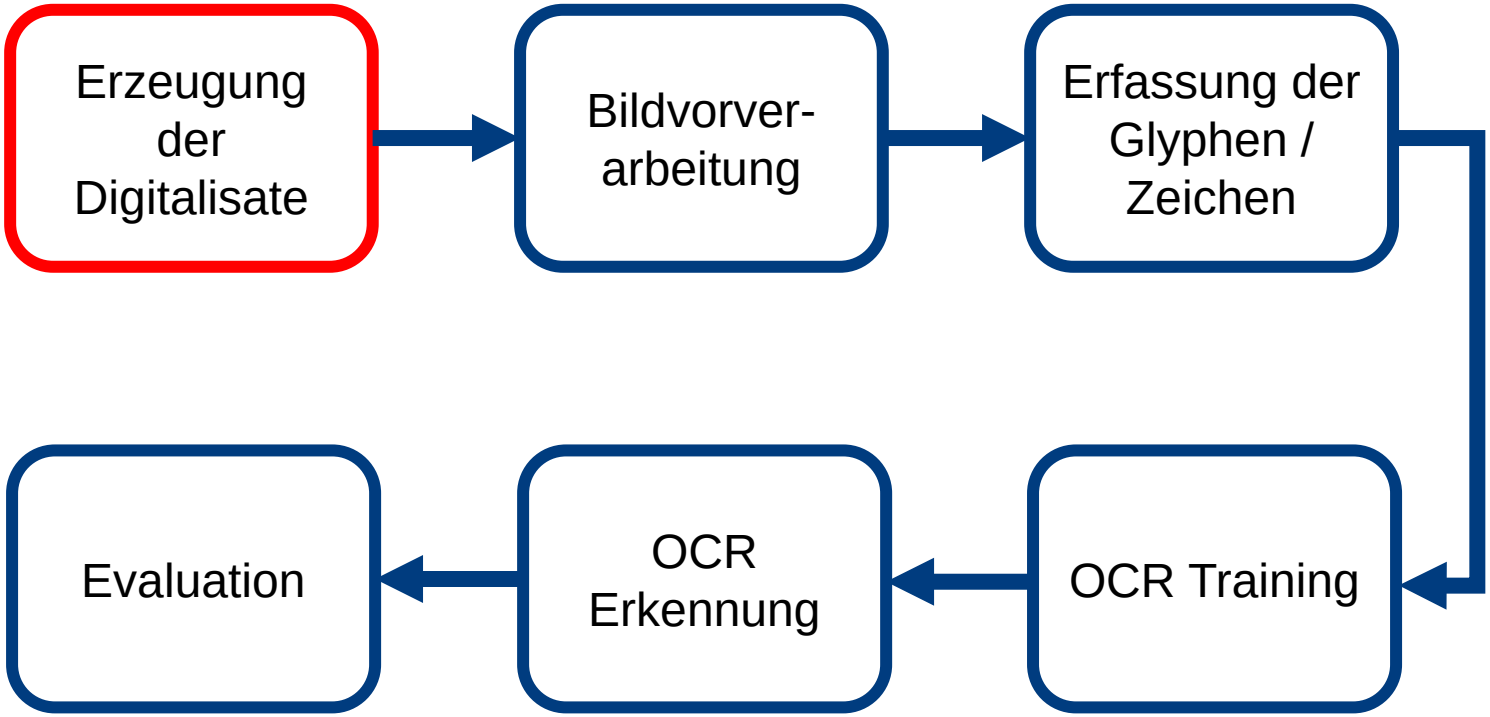
Tagesordnung

- 13:40 Digitalisate
- 14:00 Glyphen
- 14:15 Glyph Miner
- 14:45 Pause
- 15:00 Aletheia
- 15:45 Franken+
- 16:30 VietOCR

OCR Workflow



OCR Workflow

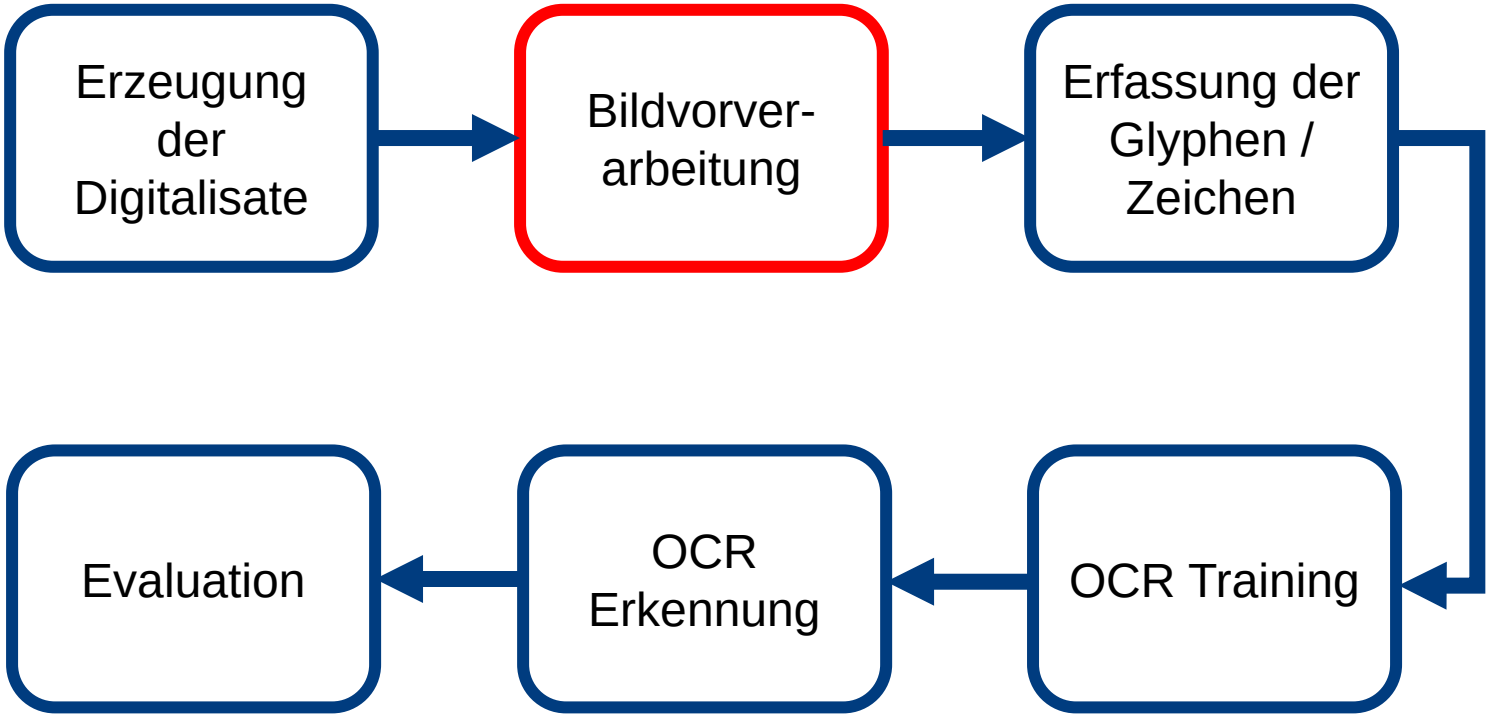


Erstellung der Digitalisate

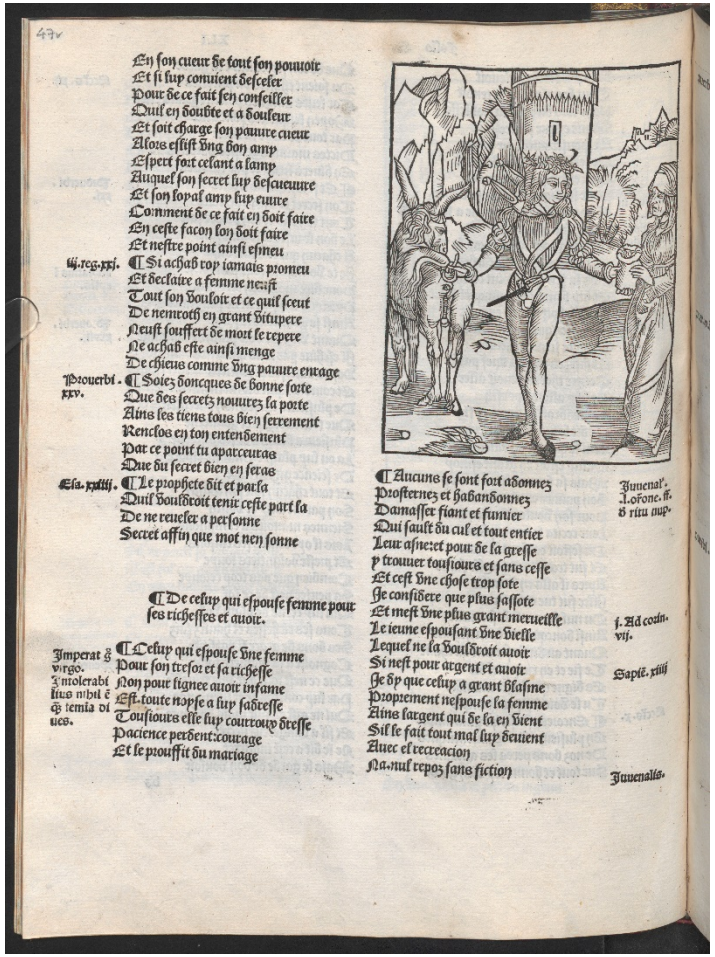
Anforderungen:

- Scanauflösung: mind. 300 DPI
- Ausgerichtete Aufnahme
- Verzerrungsfreie Aufnahme:
  - Kein Umbug im Buchfalz
  - Keine Wellen im Papier

OCR Workflow



Bildvorverarbeitung



Problemstellung:

- Automatische Entzerrung und Rotation
- Automatische Binarisierung
- Entfernung von Bildstörungen
- Segmentierung komplexer Layouts
- Beibehaltung der Lesereihenfolge

Bildvorverarbeitung: Manuell

Bildbearbeitungsprogramme,

z.B. IrfanView

- Vorteile: kostenlos, einfach zu bedienen, volle Kontrolle durch den Nutzer
- Nachteile: zeitaufwändig, keine Entzerrung und Rotation

Bildvorverarbeitung: Automatisch #1

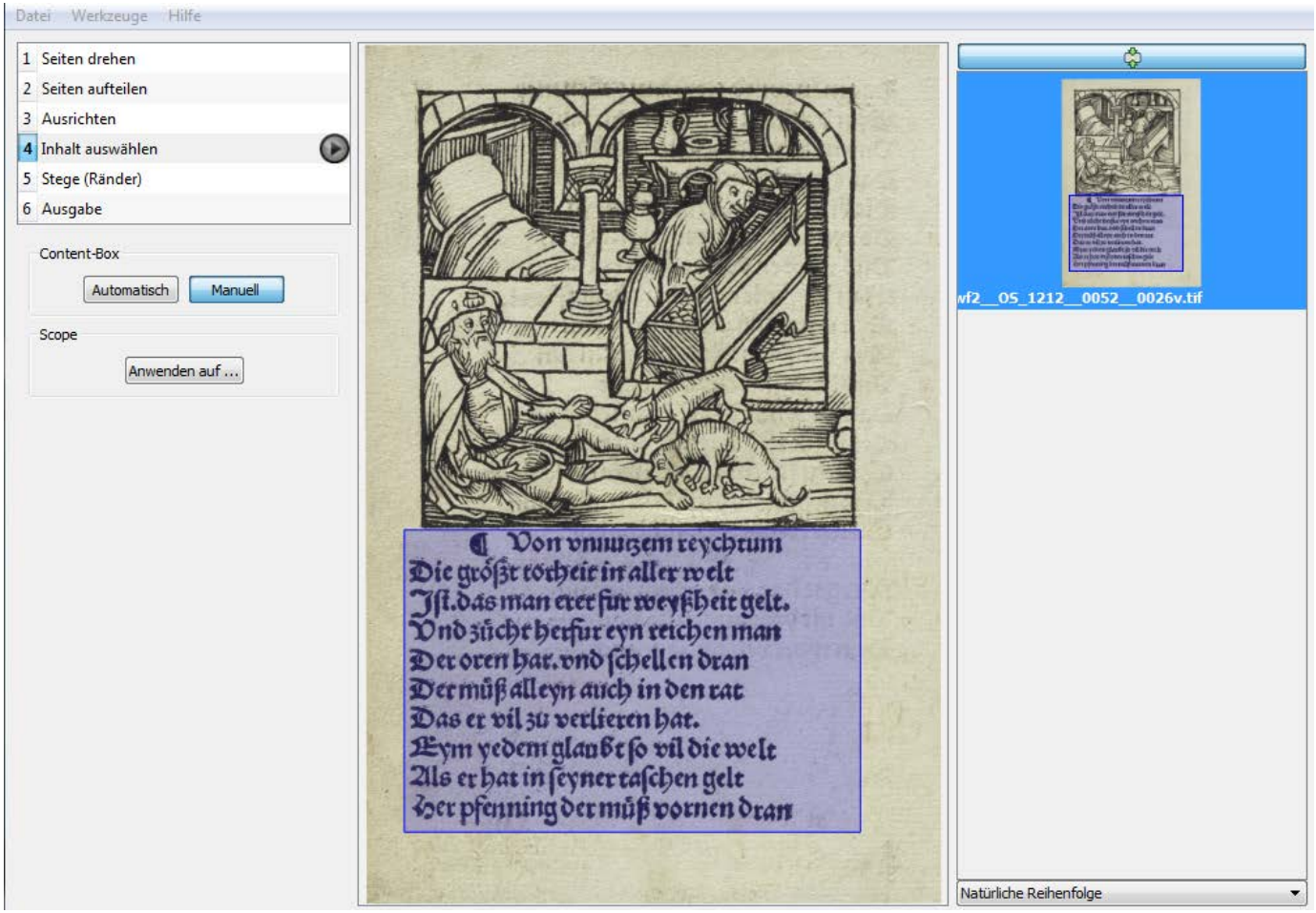
Nicht-kommerziell, z.B. ScanTailor

- Vorteile:
  - OpenSource und Plattform-unabhängig
  - halbautomatische Entzerrung, Rotation und Beschnitt
  - Binarisierung mit Entfernung von Bildstörungen wie Flecken und Rauschen

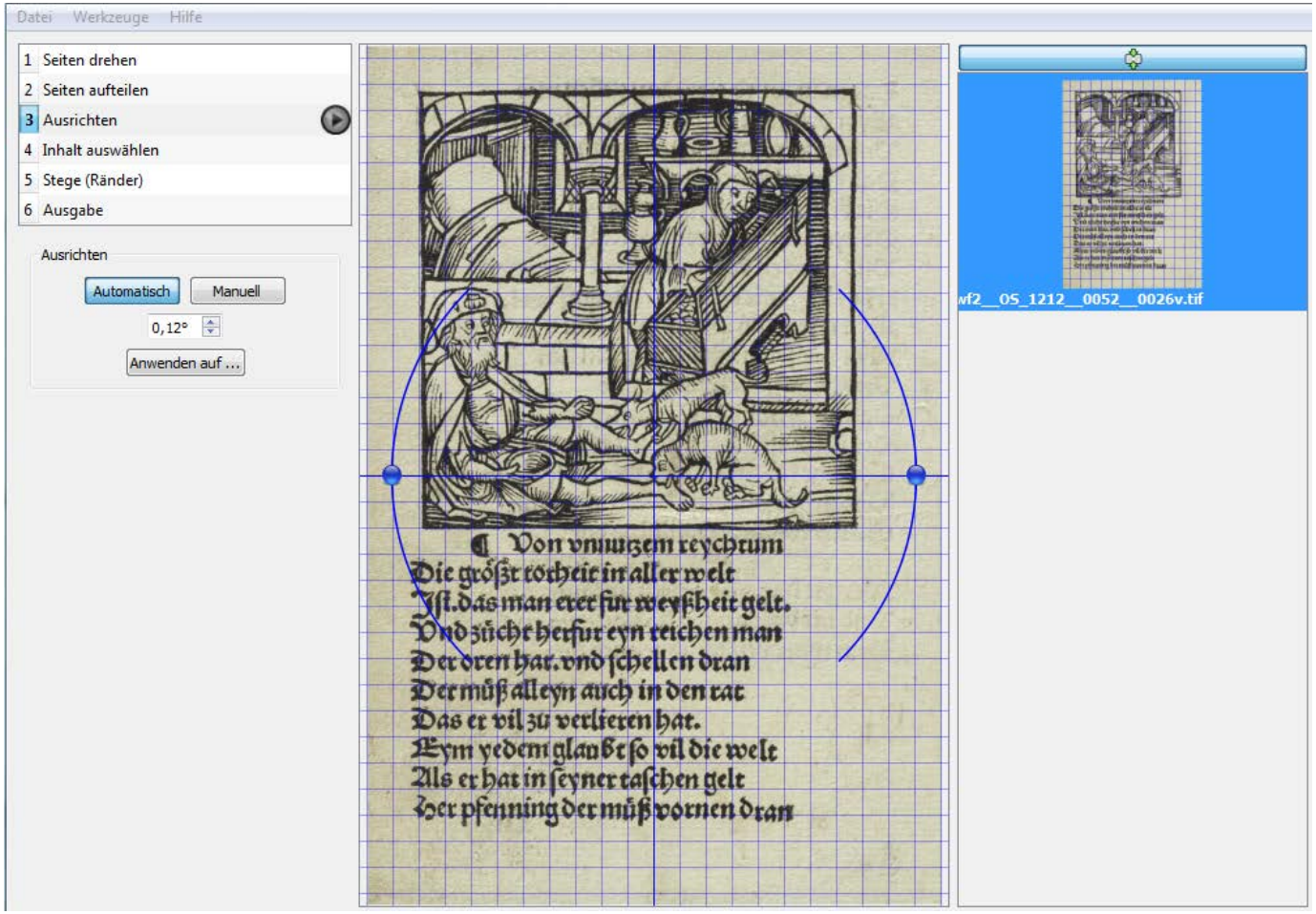
Bildvorverarbeitung: Automatisch #2

- Nachteile:
  - Automatik-Funktionen liefern teilweise falsche Ergebnisse
  - Komplexere Layouts schwer erfassbar
  - Nachkontrolle durch den Nutzer notwendig

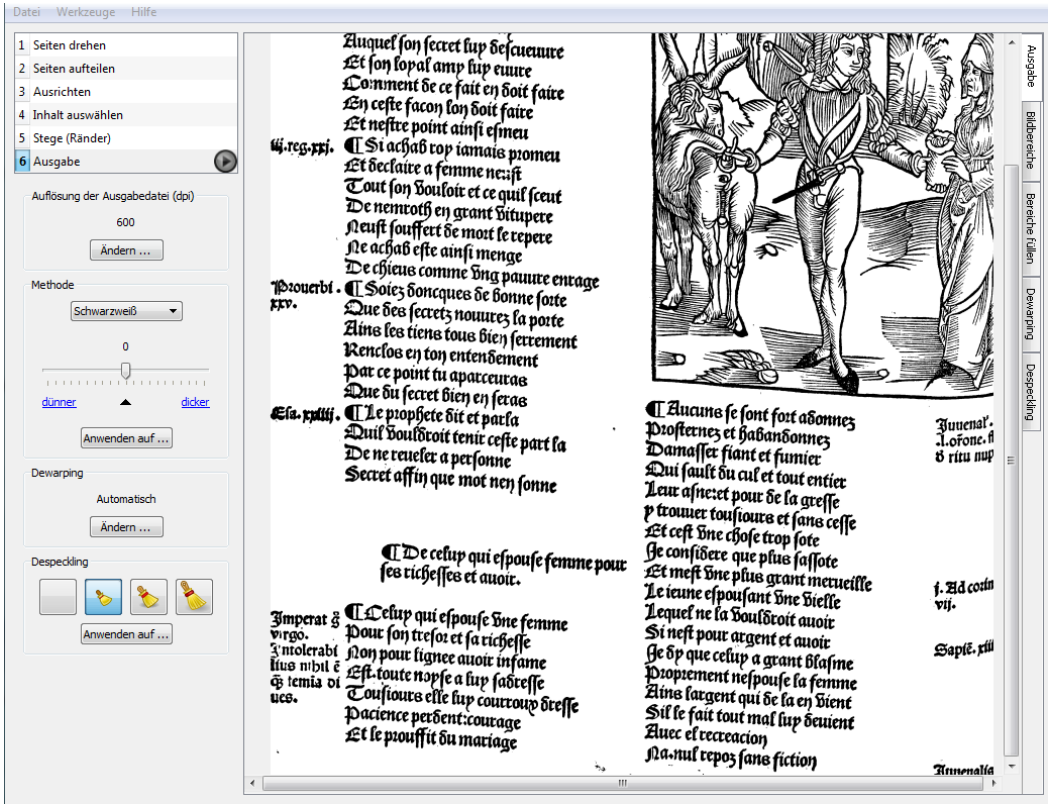
Bildvorverarbeitung: ScanTailor #1

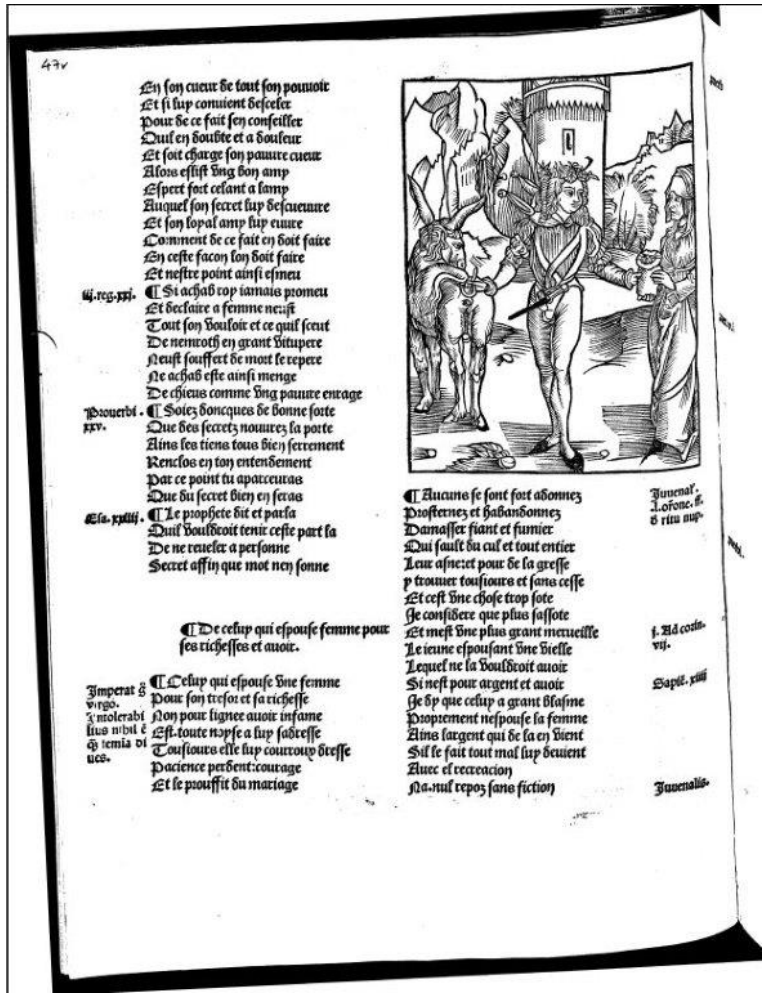


Bildvorverarbeitung: ScanTailor #2



Bildvorverarbeitung: ScanTailor vs. Abbyy FineReader



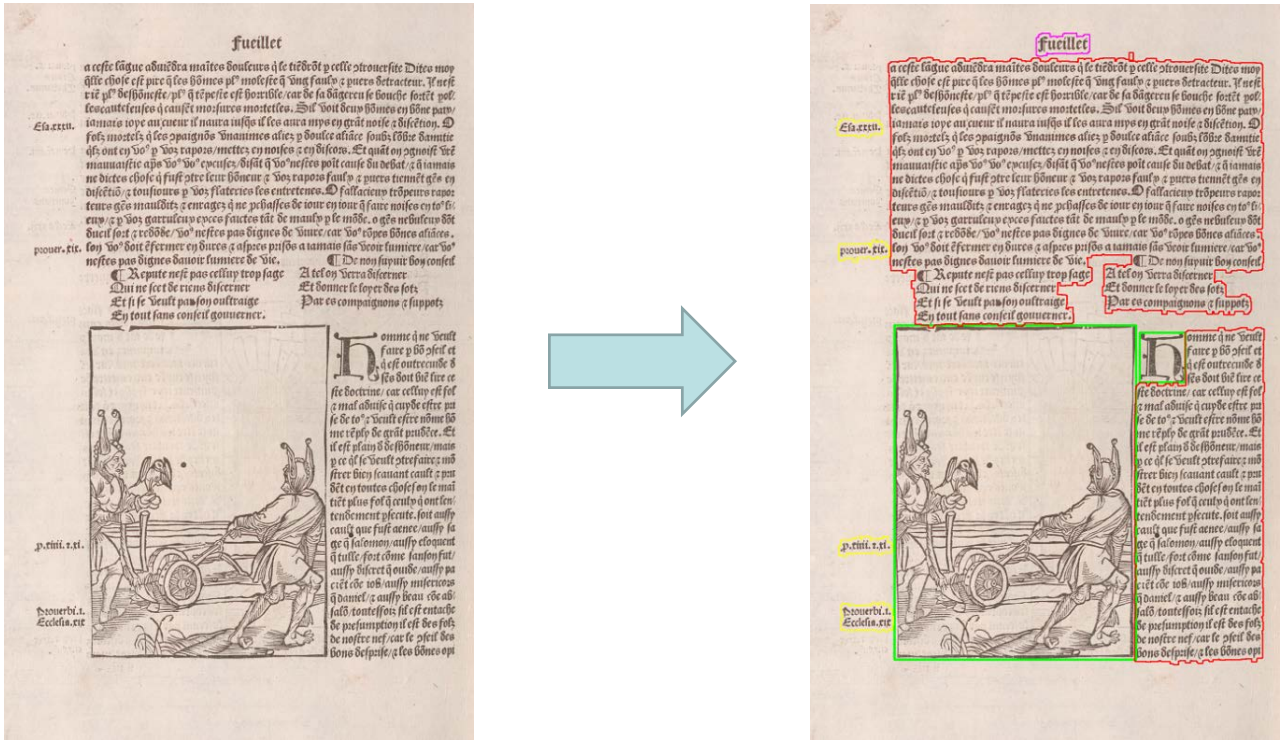


Segmentierung von Drucken

Christian Reul

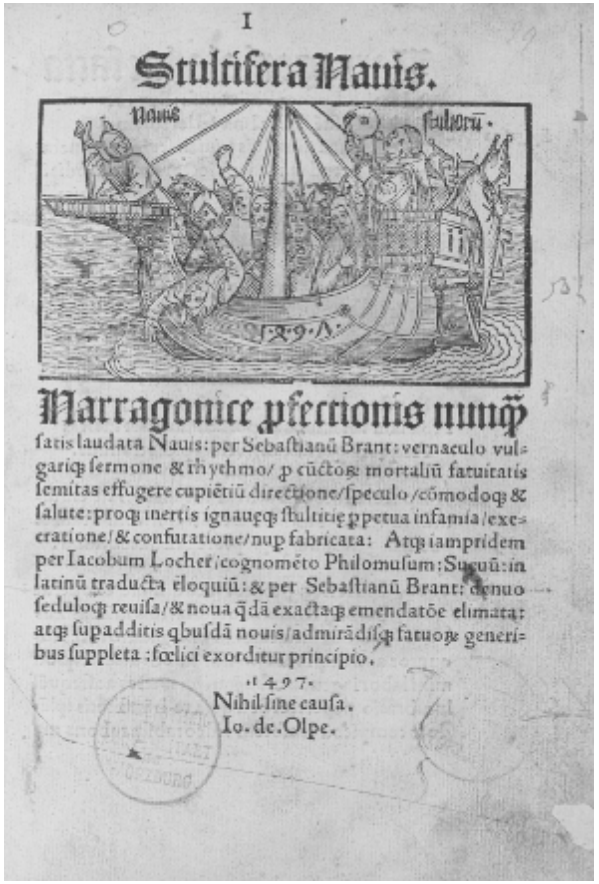
Prof. Dr. Frank Puppe

Lehrstuhl für Künstliche Intelligenz und Angewandte Informatik



Motivation und Konzept

- Einfache, schnelle Segmentierung ausgewählter Werke.
  - Verschiedene Narrenschiff-Ausgaben.
  - Der neue/praktische Schulmann.
- Anwendung grundlegender Methoden aus der Bildverarbeitung.
- Aufzeigen und Lösen/Umgehen der Probleme, die bei einem simplen Ansatz auftreten.
- Tool zur einfachen Korrektur einer fehlerhaften Lösung.


I

Stultifera Nauis.

Narragonice ꝓfectionis nunꝯ

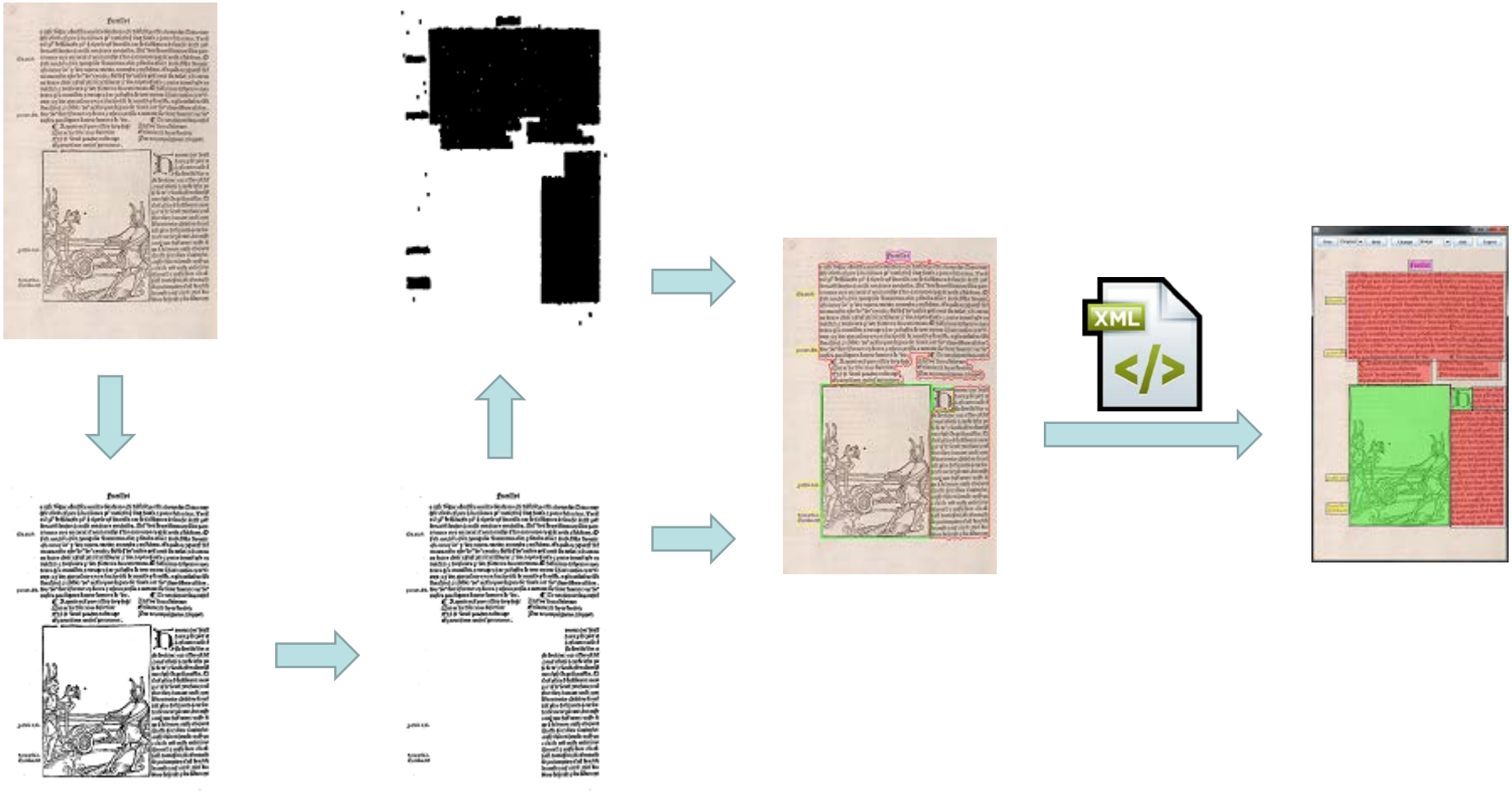
satis laudata Nauis: per Sebastianū Brant: vernaculo vulgariq; sermone & rhythmo/ p cūctos⁊ mortaliū fatuitatis semitas effugere cupiētiū directione/speculo/cōmodoq; & salute: proq; inertis ignauęq; stultitię ꝓpetua infamia/execratione/& confutatione/nup fabricata: Atq; iampridem per Iacobum Locher/cognomēto Philomusum: Sueuū: in latinū traducta eloquiū: & per Sebastianū Brant: denuo sedulocq; reuisa/& noua q̄dā exactaq; emendatōe elimata: atq; supadditis qbusdā nouis/admirādisq; fatuoꝝ generibus suppleta: fœlici exordit̄ principio.

M.cccc.xcvii.

Nihil sine causa.

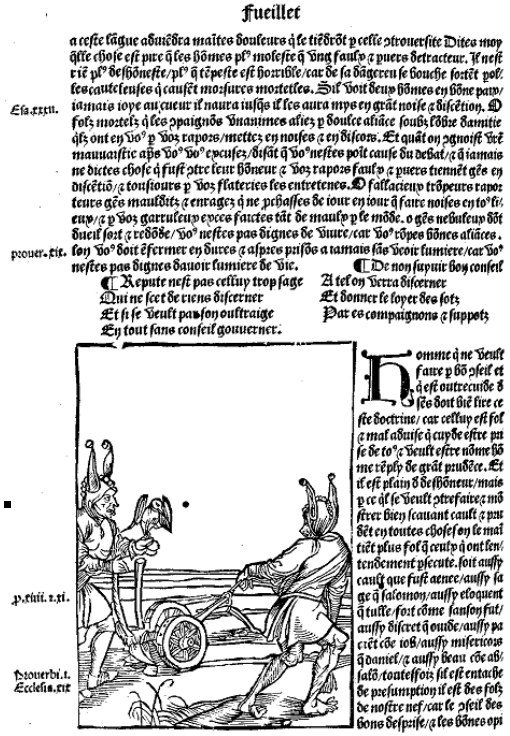
Io. de. Olpe.

Schematischer Ablauf



Vorverarbeitung und Bilddetektion

- Zuschneiden und Ausrichten (Uni-Bibliothek).
- (Temporäre) Skalierung auf einheitliche Größe.
- Binärisierung nach Otsu.
- Bilddetektion und -entfernung durch Konturfindung.



Binärdarstellung.

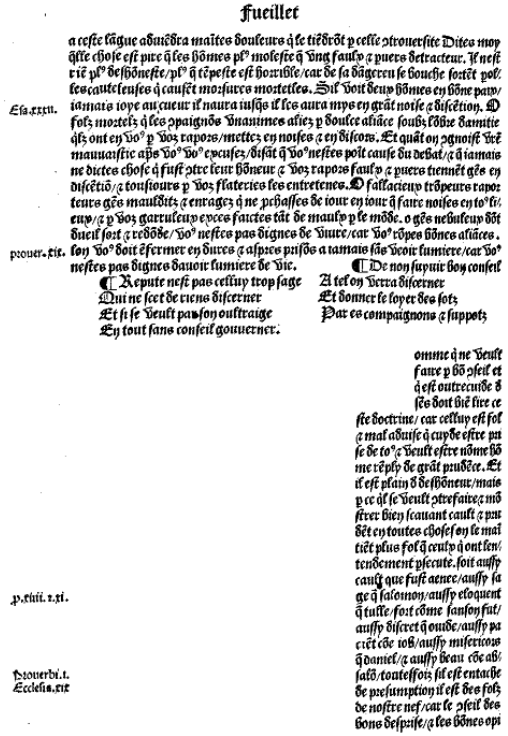
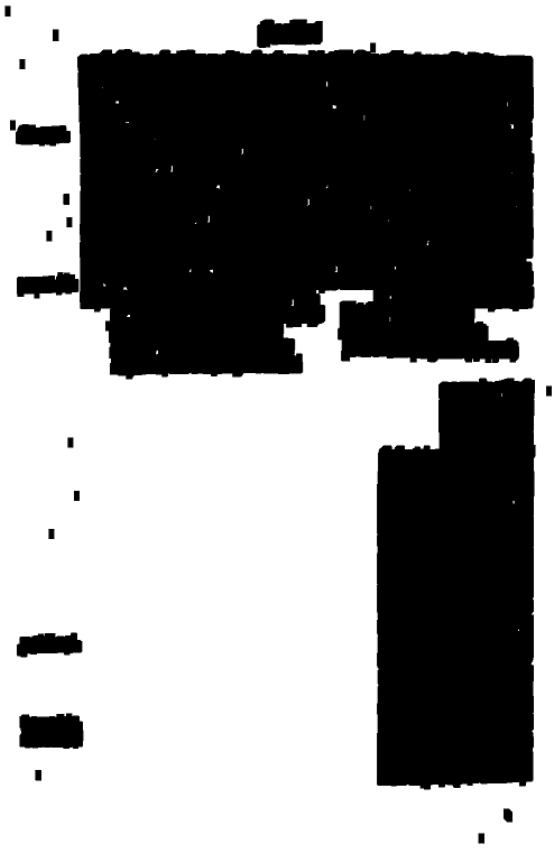


Bild und Initiale entfernt.

Klassifikation von Textabschnitten

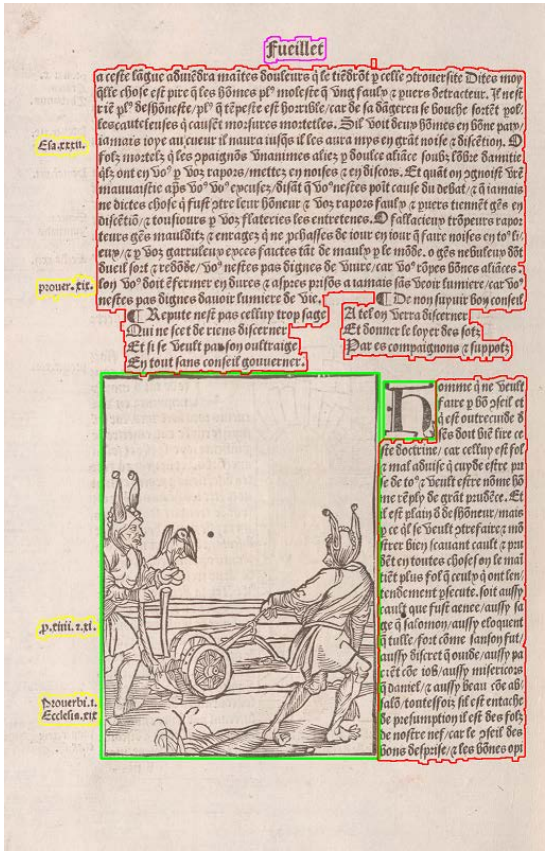
- Verbinden von Regionen durch Schließen von Zwischenräumen.
- Klassifikation anhand von Größe und Position.
- Kategorien:
  - Paragraph
  - Überschrift
  - Marginalie
  - Seitennummer
  - …



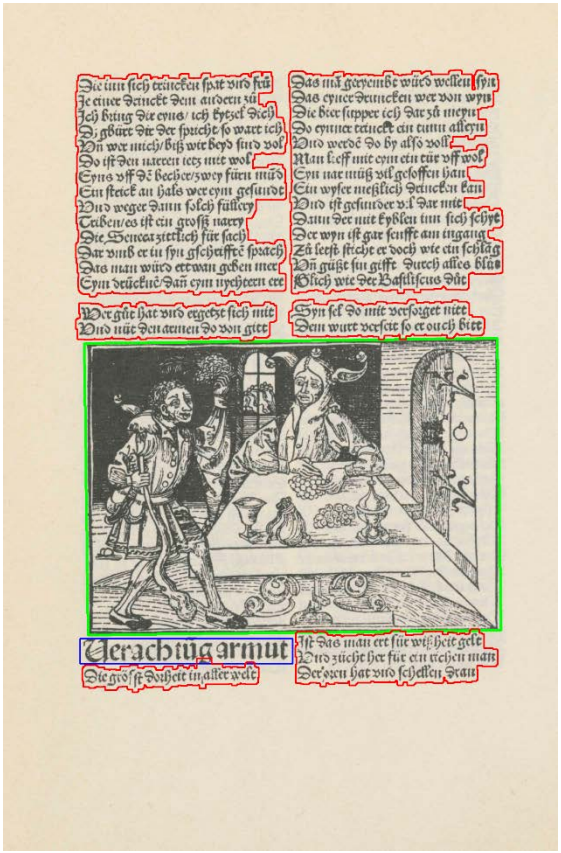
Verbundene Textregionen.

Finale Segmentierung.

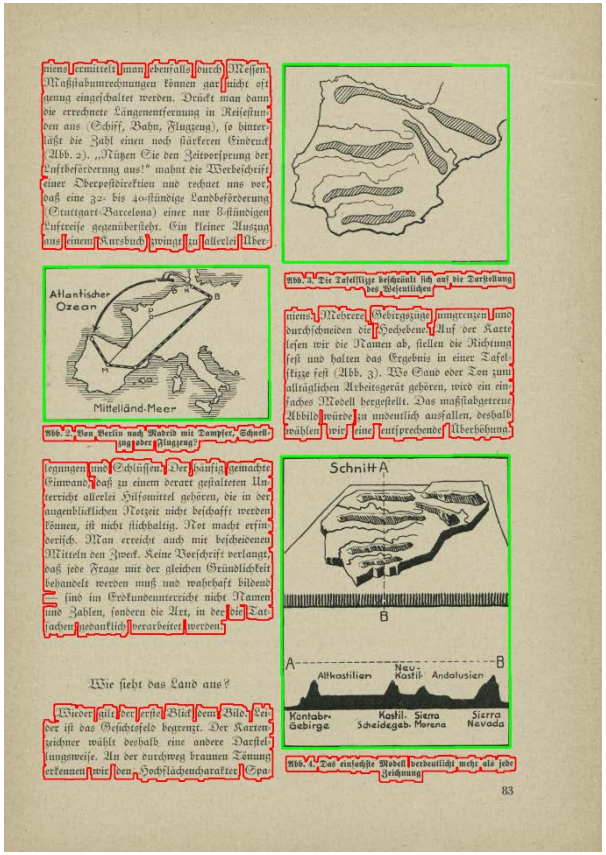
Beispielsegmentierungen



Narrenschiff GW5060, Lyon 1499.



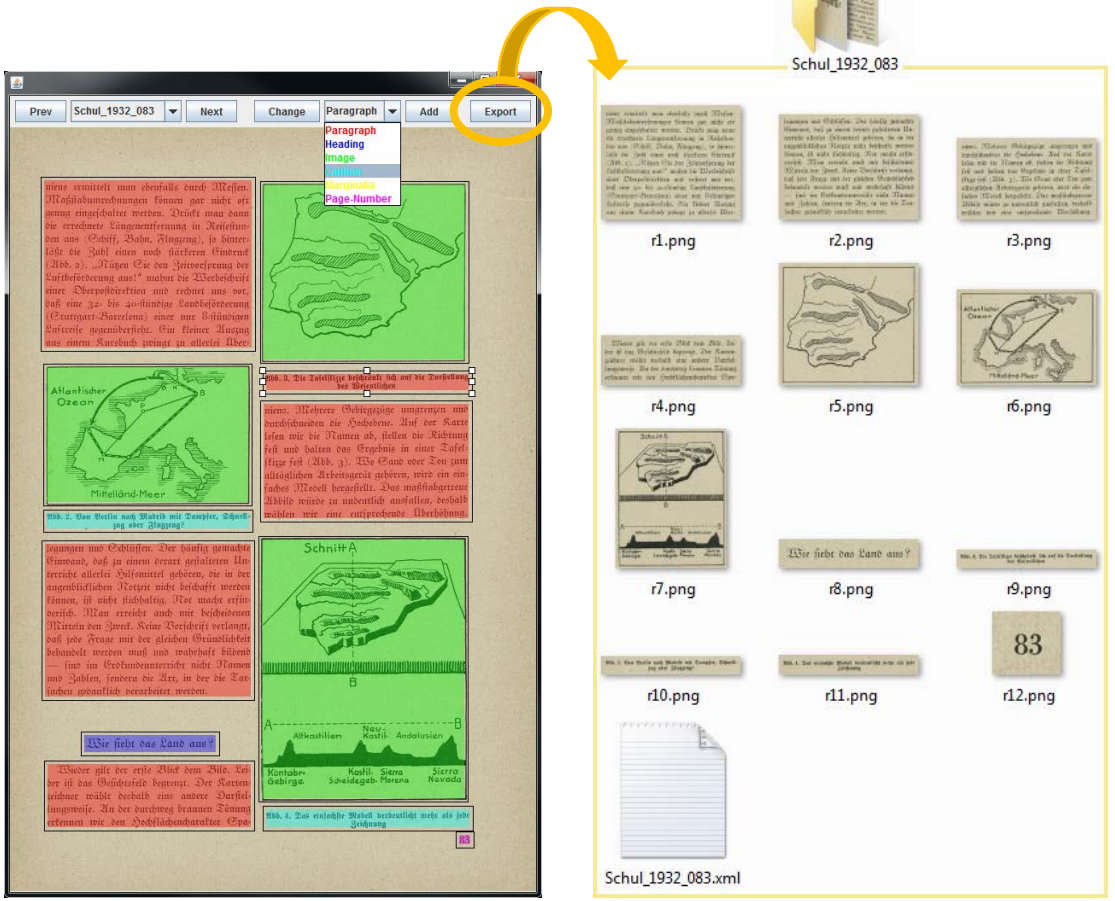
Narrenschiff GW5048, Straßburg 1494.



Der praktische Schulmann, 1932 - Heft 6.

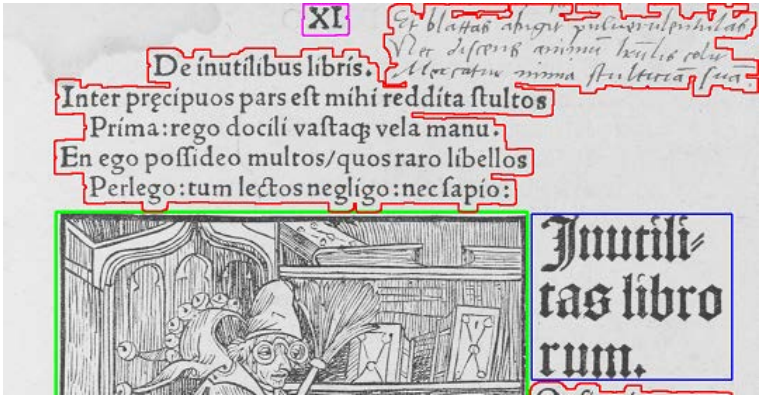
Tool zur manuellen Nachbearbeitung

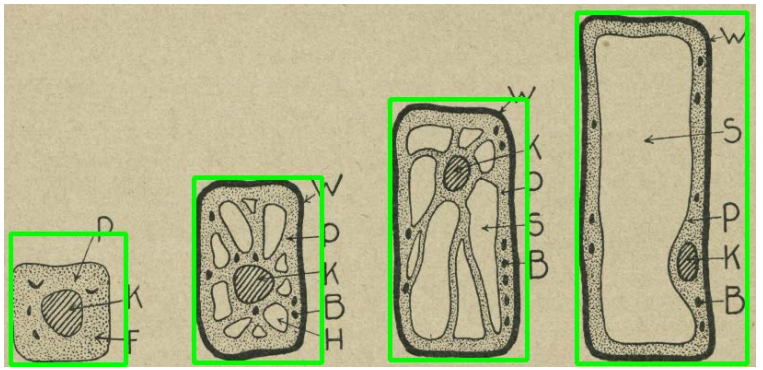
- Anzeigen des Bildes und der zugehörigen Segmentierung (PageXML).
- Löschen, Hinzufügen und Ändern von Regionen.
- Abspeichern der einzelnen Regionen und der zugehörigen PageXML-Datei.



Probleme und Lösungsansätze bei der Segmentierung von Drucken I

- Verunreinigungen der Ausgangsbilder.
  - Allgemeingültige Lösung schwierig.
  - Oft durch simple Nachbearbeitung behebbar.
- Mit diesem Ansatz keine verlässliche Bilddetektion möglich.
  - Bisher lediglich Orientierung an der äußeren Begrenzung des Bildes.
  - Fehldetektionen können die Segmentierung der gesamten Seite quasi unbrauchbar machen.





Probleme und Lösungsansätze bei der Segmentierung von Drucken II

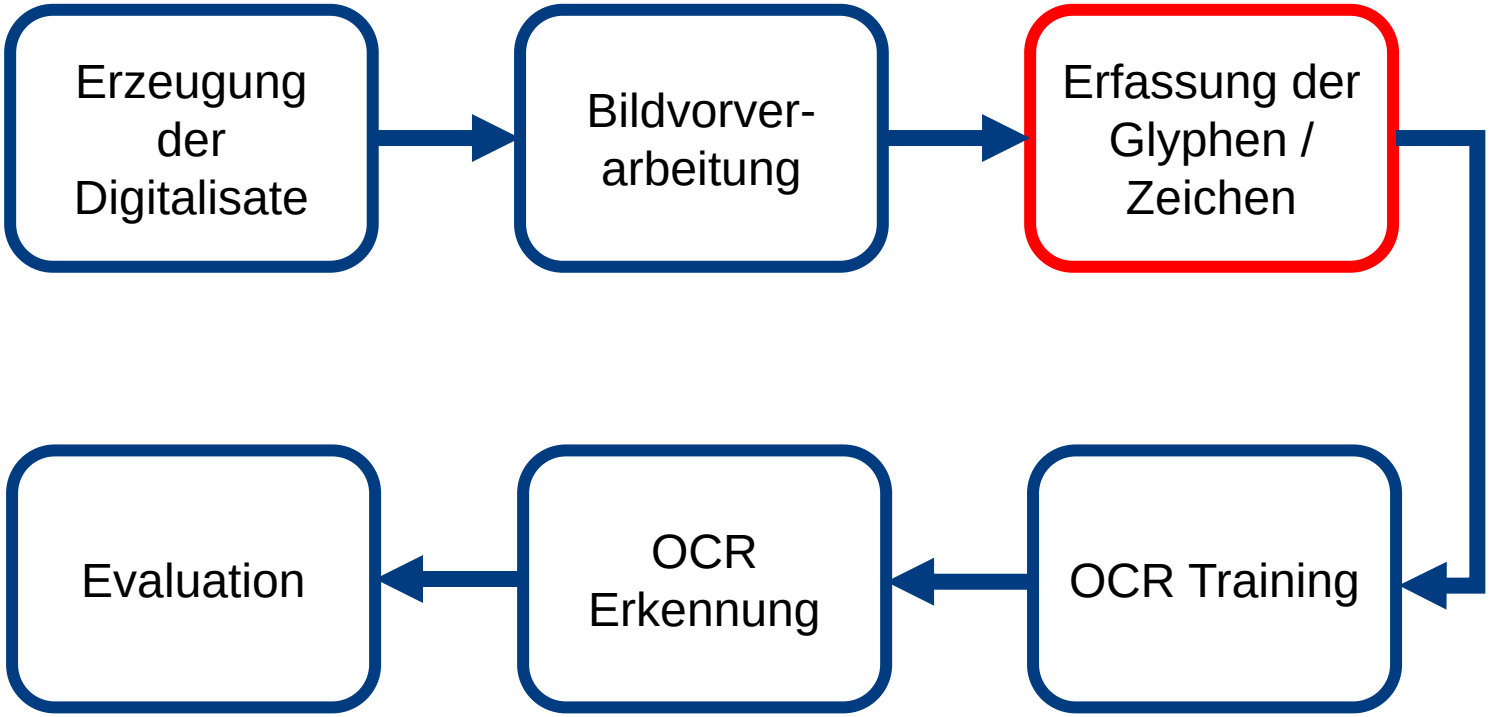
- Beschränkte Generalisierbarkeit durch starke Parametrisierung.
  - Nachteil: Manueller Aufwand.

    → Unterstützung denkbar durch:
    - (Halb)automatische Parameteroptimierung durch iterativen Ansatz.
    - Interaktive Festlegung (z. B. Konfiguration anhand von Beispielseite).
  - Vorteil: Individuell anpassbar.

Fazit:

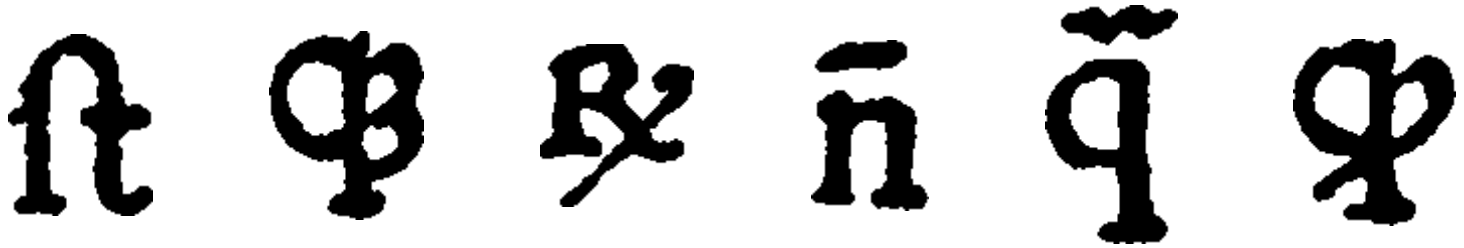
- Sehr fordernde Aufgabe, obwohl für den Menschen trivial.
- Selbst ein simpler Ansatz kann gute Ergebnisse liefern.
- Qualität stark abhängig vom zu segmentierenden Werk.
- Universallösung mit simplen Methoden unmöglich.

OCR Workflow



Glyphen #1

- Frühe Drucke beinhalten häufig eine Vielzahl an Sonderzeichen:
  - Ligaturen
  - spezielle Abkürzungen
  - Vokale mit Diakritika



Glyphen #2

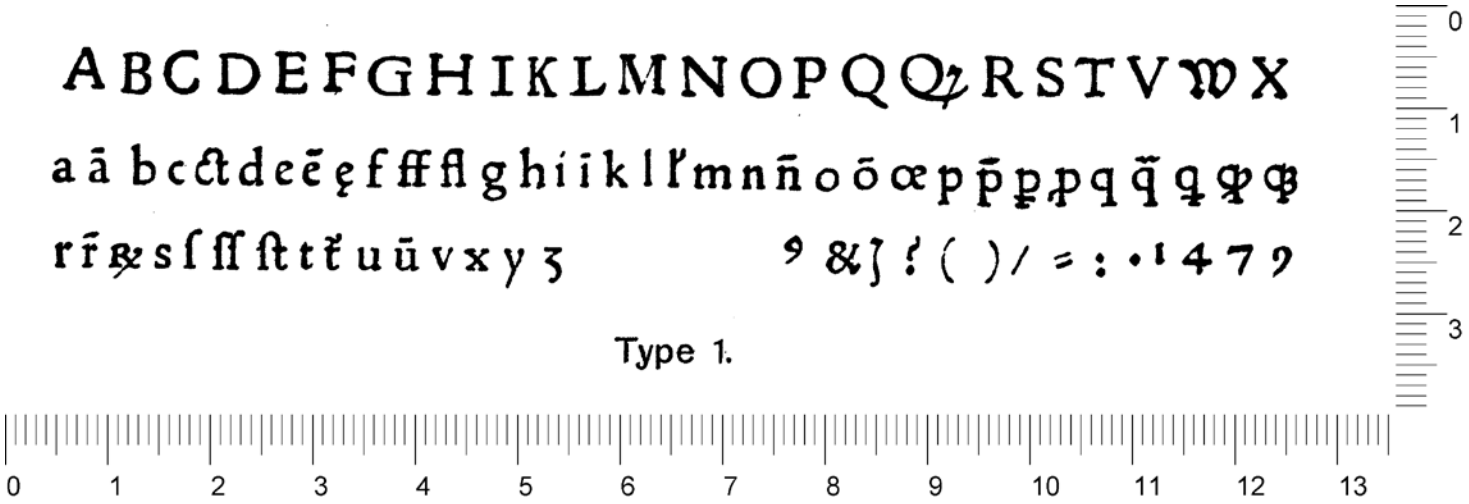
- Die Sonderzeichen sollten für eine 1:1 Übertragung bei der OCR weitestgehend erhalten bleiben:
  - Kein Vorgreifen auf eine wissenschaftliche Auswertung
  - Abkürzungen können kontextabhängig sein
  - Ligaturen können Wortgrenzen verdeutlichen

Glyphen - MUFI

- Medieval Unicode Font Initiative (MUFI):
  - Internationales Projekt zur Kodierung spezieller lateinischer Zeichen im Unicode Standard
  - Verwendet die sog. Private Use Area des Unicode-Standards

Glyphen - Typenrepertorium

- Typenrepertorium der Wiegendrucke:
  - beinhaltet die Typeninventare von ca. 2000 Offizinen (nicht vollständig)



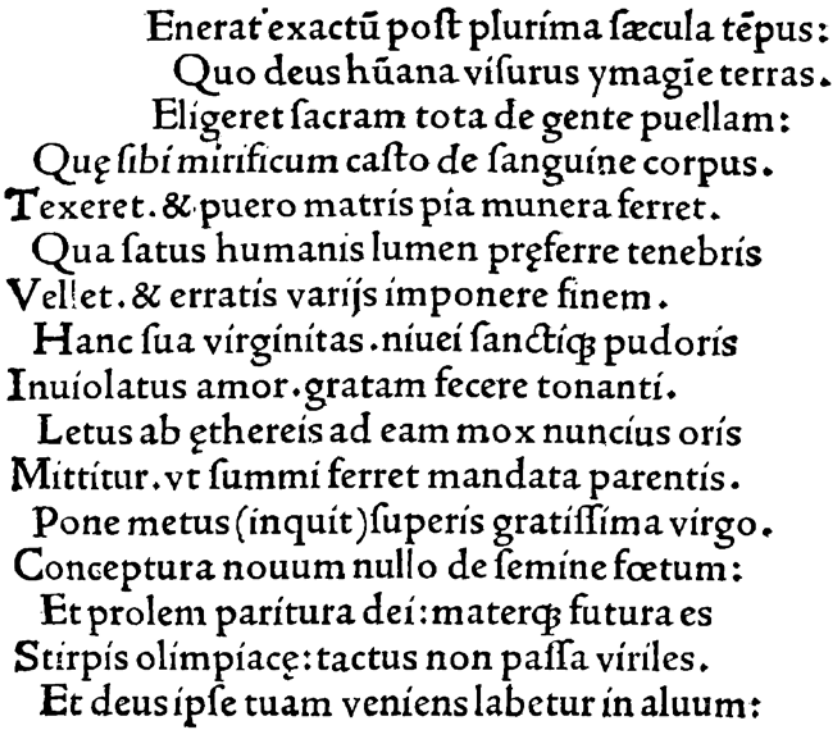
Glyphen - Typeninventar

- Ziel: Erfassung (möglichst) aller Schriftarten und Zeichen einer Offizin bzw. eines Druckers
- Beispiel: Bergmann von Olpe (Basel)
  - Type 1:109R, 2:180G, 3:77R, 4:220G, 5:109G

Glyphen – Übertragbarkeit #1

- Annahme: Wurde die OCR für alle Typen eines Druckers trainiert, können alle Werke dieser Offizin per OCR verarbeitet werden.
- Beispiel:
  - 5 von 8 Typen der Offizin Bergmann von Olpe können mit nur 2 Werken trainiert werden.

Glyphen – Übertragbarkeit #2



Enerat exactū poſt pluríma ſæcula tēpus:
Quo deus hūana víſurus ymagīe terras.
Eligeret ſacram tota de gente puellam:
Quę ſibí mirificum caſto de ſanguíne corpus.
Texeret. & puero matrís pía munera ferret.
Qua ſatus humanis lumen prę̄ferre tenebrís
Vellet. & erratís varijs ímponere finem.
Hanc ſua vírgínitas. níuei ſanctiq; pudorís
Inuíolatus amor. gratam fecere tonantí.
Letus ab ętherеís ad eam mox nuncíus orís
Mittítur. vt ſummí ferret mandata parentís.
Pone metus (inquit) ſuperís gratíſſíma vírgo.
Conceptura nouum nullo de ſemíne fœtum:
Et prolem parítura dei: materq; futura es
Stírpís olímpíacę: tactus non paſſa víríles.
Et deus ípſe tuam veníens labetur ín aluum:

Eneratoexactū poft pIurima ſæcula tēpus :
Quo deus hūanaviſurus ymagie terras .
Eligeret ſacram tota de gente puellam:
Quę ſlbi miriflcum caſto de ſanguine corpus .
Texeret. &.puero matris pia munera ferret.
Qua ſatus humanis lumen prę̨ferre tenebris
Velget. & erratis varilis imponere finem .
Hanc ſua virginitas .niuei ſanctiq; pudoris
Inuiolatus amor.gratam fecere tonanti.
Letus ab ęthereis ad eam mox nuncius oris
Mittitur.vt ſummi ferret mandata parentis .
pone metus (inquit)ſuperis gratiſſima virgo .
Conceptura nouum nullo de ſemine fœtum:
Et prolem paritura dei : materq; futura es
Stū.rpis olimpiacę: tactus non paffa viriles.
Et deus ipſe tuam veniens labetur in aluum:

Glyph Miner

- Ziel: Effiziente Erfassung von Glypheninventaren (mit Beispielen)

- Entwickelt an der Uni Würzburg
  https://github.com/benedikt-budig/glyph-miner

Aletheia

- Entwickelt von PRImA *(Pattern Recognition & Image Analysis Research Lab)*
- University of Salford, Manchester
- http://www.primaresearch.org/tools/Aletheia

Aletheia

- Ziel: PAGE-XML (2010)

```
<Page imageFilename="GW5042__DE-20_FLDE-Swf2__OS_1212__0024__0011v_bw.tif" imageWidth="2038" imageHeight="3025">
  <TextRegion id="r1" type="paragraph">
    <Coords>
    <TextLine id="l2">
      <Coords>
      <Word id="w3">
        <Coords>
        <Glyph id="c4">
          <Coords>
            <Point x="315" y="200"/>
            <Point x="315" y="285"/>
            <Point x="417" y="285"/>
            <Point x="416" y="200"/>
          </Coords>
          <TextEquiv>
            <PlainText/>
            <Unicode>G</Unicode>
          </TextEquiv>
        </Glyph>
        <Glyph id="c5">
          <Coords>
            <Point x="420" y="219"/>
            <Point x="420" y="275"/>
            <Point x="456" y="275"/>
            <Point x="456" y="219"/>
```

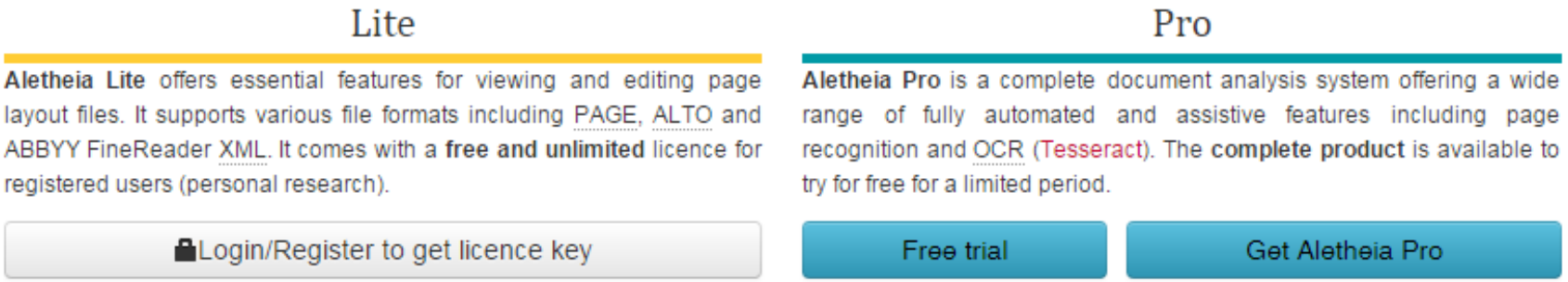
Aletheia

Lite

- ~~Binarisierung~~
- ~~Schmutzentfernung~~
- Manuelle Zuweisung

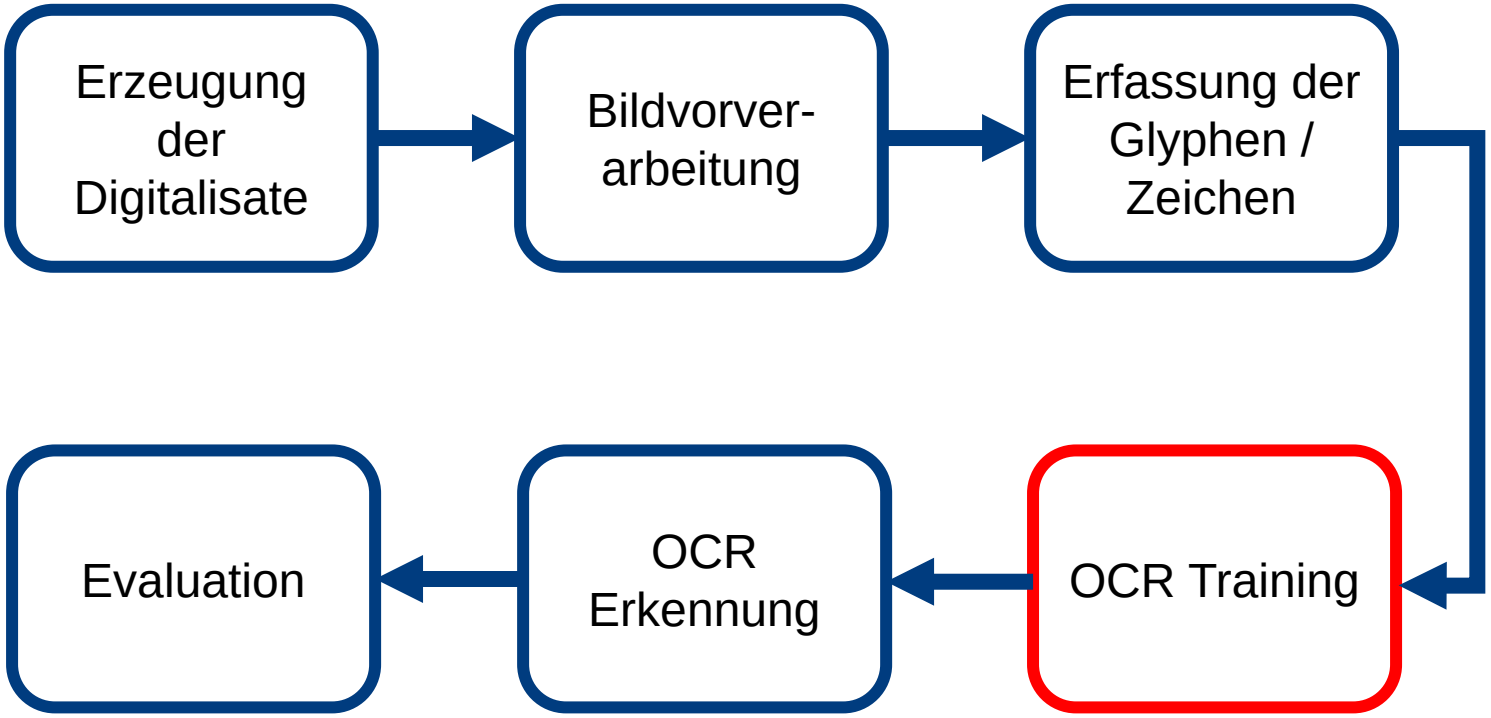
Pro

- Binarisierung
- Schmutzentfernung
- Automatische Erkennung (Regionen, Zeilen, Wörtern, Glyphen)



http://www.primaresearch.org/tools/Aletheia/Editions

OCR Workflow



Franken+

- Entwickelt vom eMOP *(Early Modern OCR Project)*
- Texas A&M University
- http://emop.tamu.edu/outcomes/Franken-Plus

Voraussetzungen:

- .NET Framework
- MySQL
- Tesseract

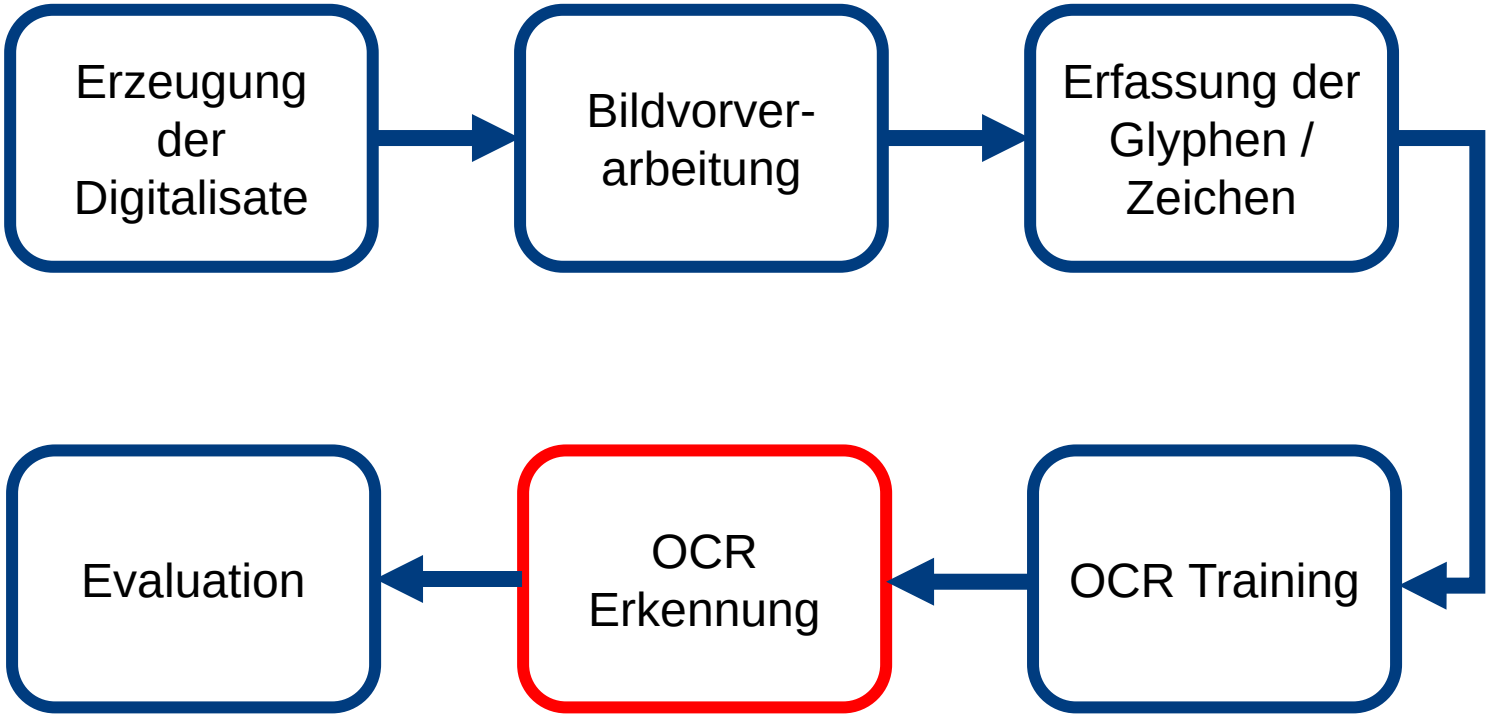
Material:

- PAGEXML
- *Ground Truth* in Textform

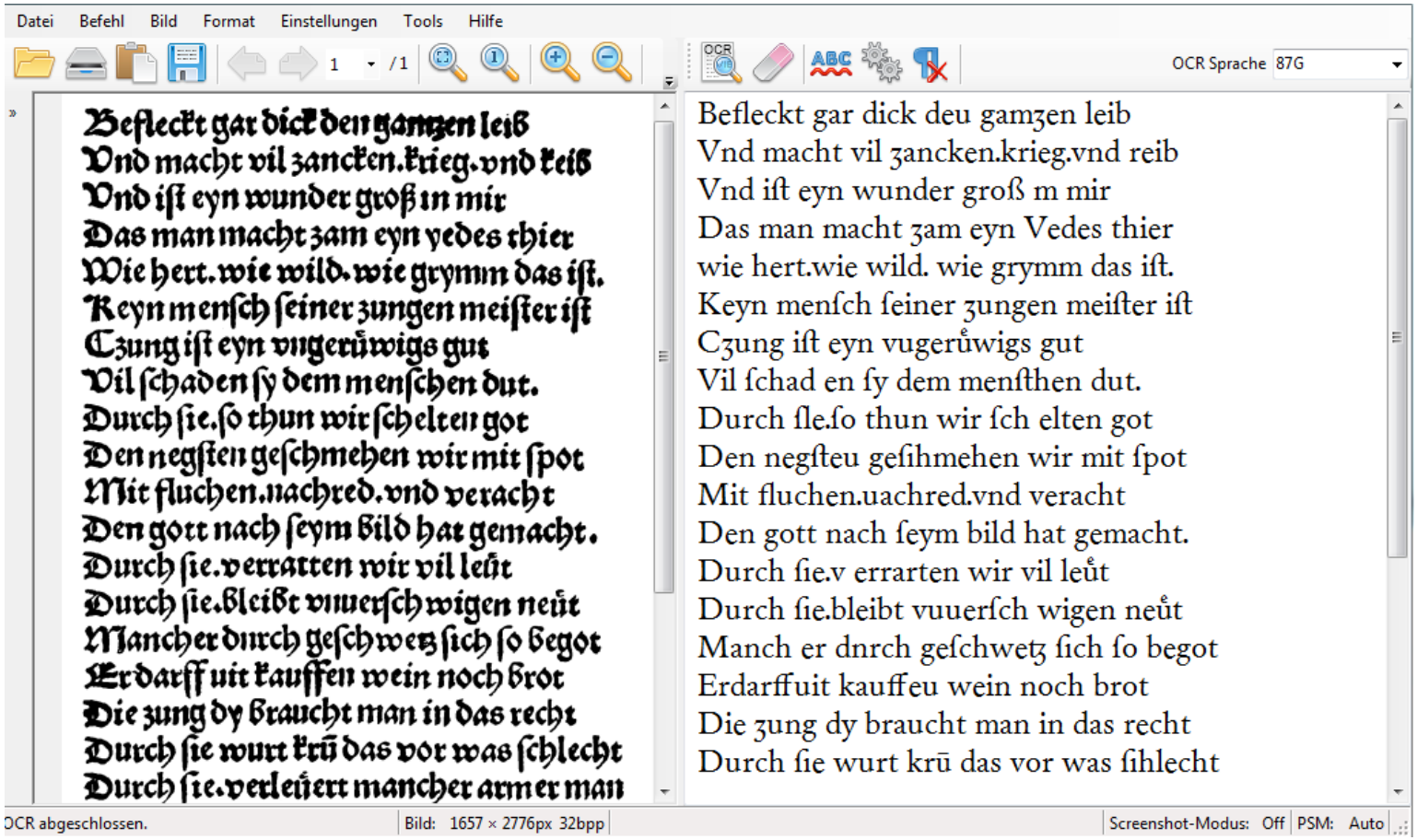
Arbeitsschritte Franken+

1. Font und Language erstellen
2. Glyphen (PAGEXML) importieren
3. Font bearbeiten
4. Synthetischen Text erstellen
5. Language trainieren

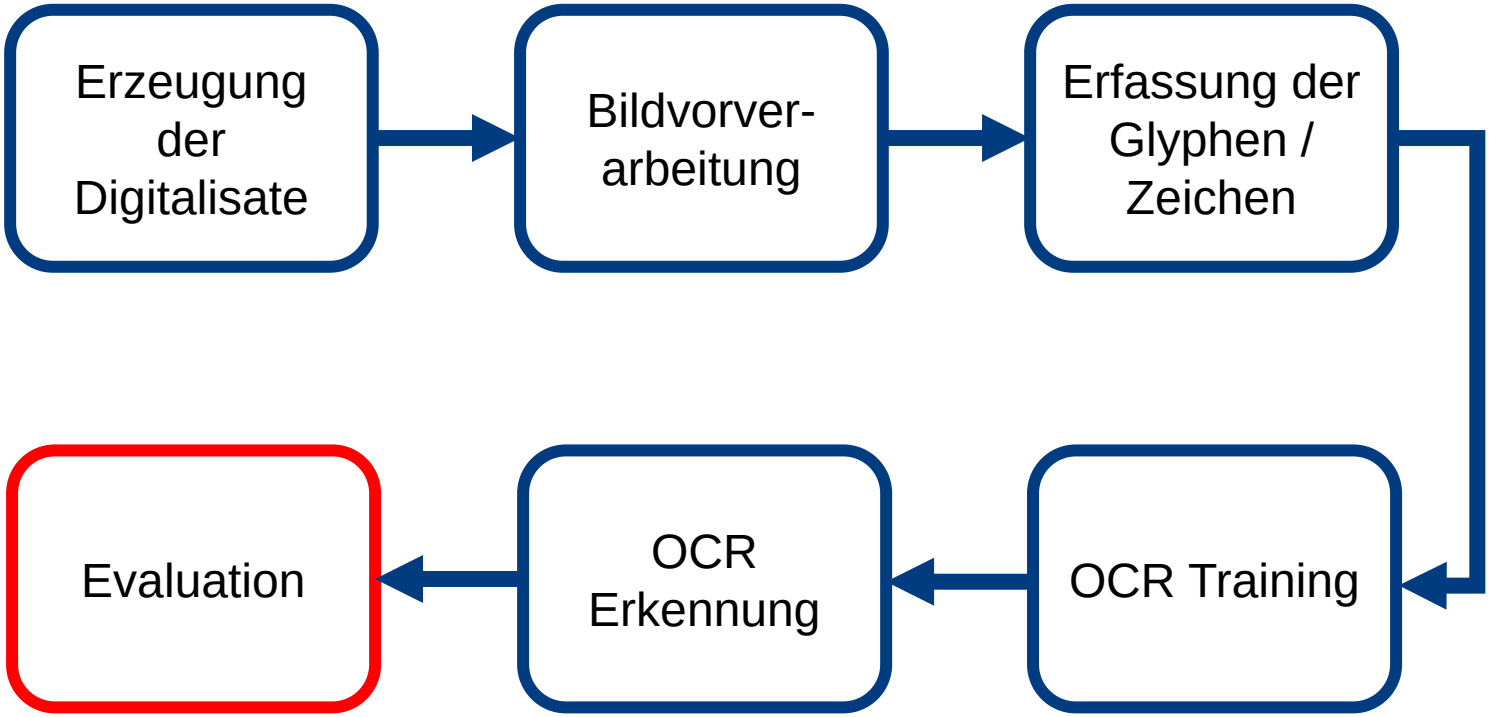
OCR Workflow



VietOCR


Datei Befehl Bild Format Einstellungen Tools Hilfe
OCR Sprache 87G
Befleckt gar dick den gantzen leib
Vnd macht vil zancken.krieg.vnd leib
Vnd iſt eyn wunder groß in mir
Das man macht zam eyn vedes thier
Wie hert.wie wild.wie grymm das iſt.
Keyn menſch ſeiner zungen meiſter iſt
Czung iſt eyn vngerůwigs gut
Vil ſchaden ſy dem menſchen dut.
Durch ſie.ſo thun wir ſchelten got
Den negſten geſchmehen wir mit ſpot
Mit fluchen.nachred.vnd verachț
Den gott nach ſeym bild hat gemachț.
Durch ſie.verratten wir vil leůt
Durch ſie.bleibt vnuerſch wigen neůt
Mancher dnrch geſchwetz ſich ſo begot
Erdarff uit kauffen wein noch brot
Die zung dy braucht man in das recht
Durch ſie wurt krū das vor was ſchlecht
Durch ſie.verleůert mancher armer man
Befleckt gar dick deu gamzen leib
Vnd macht vil zancken.krieg.vnd reib
Vnd iſt eyn wunder groß m mir
Das man macht zam eyn Vedes thier
wie hert.wie wild. wie grymm das iſt.
Keyn menſch ſeiner zungen meiſter iſt
Czung iſt eyn vugerůwigs gut
Vil ſchad en ſy dem menſthen dut.
Durch ſle.ſo thun wir ſch elten got
Den negſteu geſihmehen wir mit ſpot
Mit fluchen.uachred.vnd veracht
Den gott nach ſeym bild hat gemacht.
Durch ſie.v errarten wir vil leůt
Durch ſie.bleibt vuuerſch wigen neůt
Manch er dnrch geſchwetz ſich ſo begot
Erdarffuit kauffeu wein noch brot
Die zung dy braucht man in das recht
Durch ſie wurt krū das vor was ſihlecht
OCR abgeschlossen.
Bild: 1657 × 2776px 32bpp
Screenshot-Modus: Off PSM: Auto

OCR Workflow



TextEvaluation