



# Perspektiven der automatischen Texterfassung als Grundlage wissenschaftlicher Editionen

am Beispiel der Brief- und Schriftenausgabe  
der Bernd Alois Zimmermann-Gesamtausgabe

Matthias Boenig & Kay-Michael Würzner  
[{boenig|wuerzner}@bbaw.de](mailto:{boenig|wuerzner}@bbaw.de)

Geisteswissenschaftliche Forschungsdaten.  
Methoden zur digitalen Erfassung, Aufbereitung und Präsentation  
19. Oktober 2017



# Übersicht

- Einleitung
  - ▶ die BERND ALOIS ZIMMERMANN-Gesamtausgabe
  - ▶ automatische Texterfassung
  - ▶ Motivation
- Workflowbeschreibung
  - ▶ Bildvorverarbeitung
  - ▶ Layoutanalyse
  - ▶ Zeichenerkennung
  - ▶ Textbearbeitung
- Perspektiven
  - ▶ Volltextverbesserung
  - ▶ OCR-D
  - ▶ Editionsunterstützung





# Einleitung



# Die Bernd Alois Zimmermann-Gesamtausgabe

- gemeinsames **Langzeitvorhaben** der Akademien in Mainz und Berlin
- jüngste musikwissenschaftliche Gesamtausgabe
- Ziele:
  - ▶ Edition der Kompositionen
  - ▶ Herausgabe der **Schriften** und **repräsentativer Teile** der Korrespondenz („Spiegel der breiteren Debattengeschichte der zweiten Hälfte des 20. Jh.“)
- Anwendung computergestützter Methoden bei **Erfassung** und Auswertung
- bisher ca. 6000 Objektseiten in der Abteilung Schriften und Briefe digitalisiert



# Die Bernd Alois Zimmermann-Gesamtausgabe

## Beispiele:

Erfüllung des neugewonnenen und neueroberten "Tonmaterials" hin, die die Wendung zur "Klassik" einerseits als stilistisch historische Synthese und ~~andererseits~~ zur weltanschaulich oder gar politischen Orientierung (Russland) als gewissermassen kosmischer Synthese<sup>andererseits</sup> <sup>in ersterer Hinsicht</sup> ~~greiflich~~ macht. In Frankreich ist es ~~einerseits~~ der emigrierte

Sie wissen, allervere  
ahren ein ausgesprochen  
stände es zwar bisher  
rung etwas einseitig b  
an, dass unsere Verbin

ken exponiert; darin  
und das Publikum  
n den Ordnungszahlen  
er gleiche Teile ge-  
de Markierung

Lieber Freund!

e wissen, um eines der sogen  
"Los alimentos del hombre"  
les zur Verherrlichung der  
bindung von Wort und Musik



# Die Bernd Alois Zimmermann-Gesamtausgabe

- **Pilotstudie** zum Einsatz von automatischer Texterfassung zur Editions Vorbereitung
- Kooperation der BAZ-GA mit OCR-D (DFG-Koordinationsprojekt zur Verbesserung von OCR-Verfahren)
- Fragestellungen:
  1. Vorteile grundständig **manueller Erfassung vs. automatischer Erfassung** mit anschließender Expertenbearbeitung
  2. **Einfluss** der oportunistischen Texterfassung (vs. vorausgehender gezielter Textauswahl) **auf den Editionsprozess**





# Automatische Texterfassung

- Textquellen mehrheitlich nicht in digitaler Form verfügbar
  - historische Bücher, Zeitungen und andere Druckerzeugnisse
  - handgeschriebene Manuskripte und Briefe
- Bewahrung/Konservierung durch Scan oder Foto
  - Bilder sind **kein Text**
  - ⇒ Textsuche oder quantitative Auswertung nicht möglich
- Automatische Erfassung von **Text** in Bildern (aka. Optical Character Recognition)
- Automatische Erfassung des **Layouts** in Bildern (aka. Optical Layout Recognition)





# Workflowbeschreibung





# Übersicht

- Ziel: Definition und Implementierung eines Workflows
  - ▶ möglichst **OpenSource**
  - ▶ möglichst ohne Programmierkenntnisse einsetzbar
  - ▶ **lokal** installierbar
  - ▶ Textergebnis als **Grundlage einer Edition** verwendbar
- Komponenten
  - ▶ Bildvorverarbeitung
  - ▶ Layoutanalyse
  - ▶ Zeichenerkennung
  - ▶ Textbearbeitung



# Übersicht



- Prozesse zur bestmöglichen Vorbereitung der Digitalisate für OLR und OCR
  - ▶ **Cropping:** Beschneidung des Digitalisats auf den Druckbereich
  - ▶ **Deskewing:** Rotation des Digitalisats zur Begradigung von Schrägstellungen
  - ▶ **Binarization:** Binäre Kodierung der Pixel (bedruckte Bereiche schwarz, nicht-bedruckte Bereiche weiß)
  - ▶ **Despeckling:** Entfernung von Bildartefakten (Verschmutzungen, sichtbare Papiermaserung etc.)
  - ▶ **Dewarping:** Begradigung von Wellen auf Zeilenebene
- starker Einfluss auf die Erkennungsqualität
- besondere Relevanz für historische Vorlagen

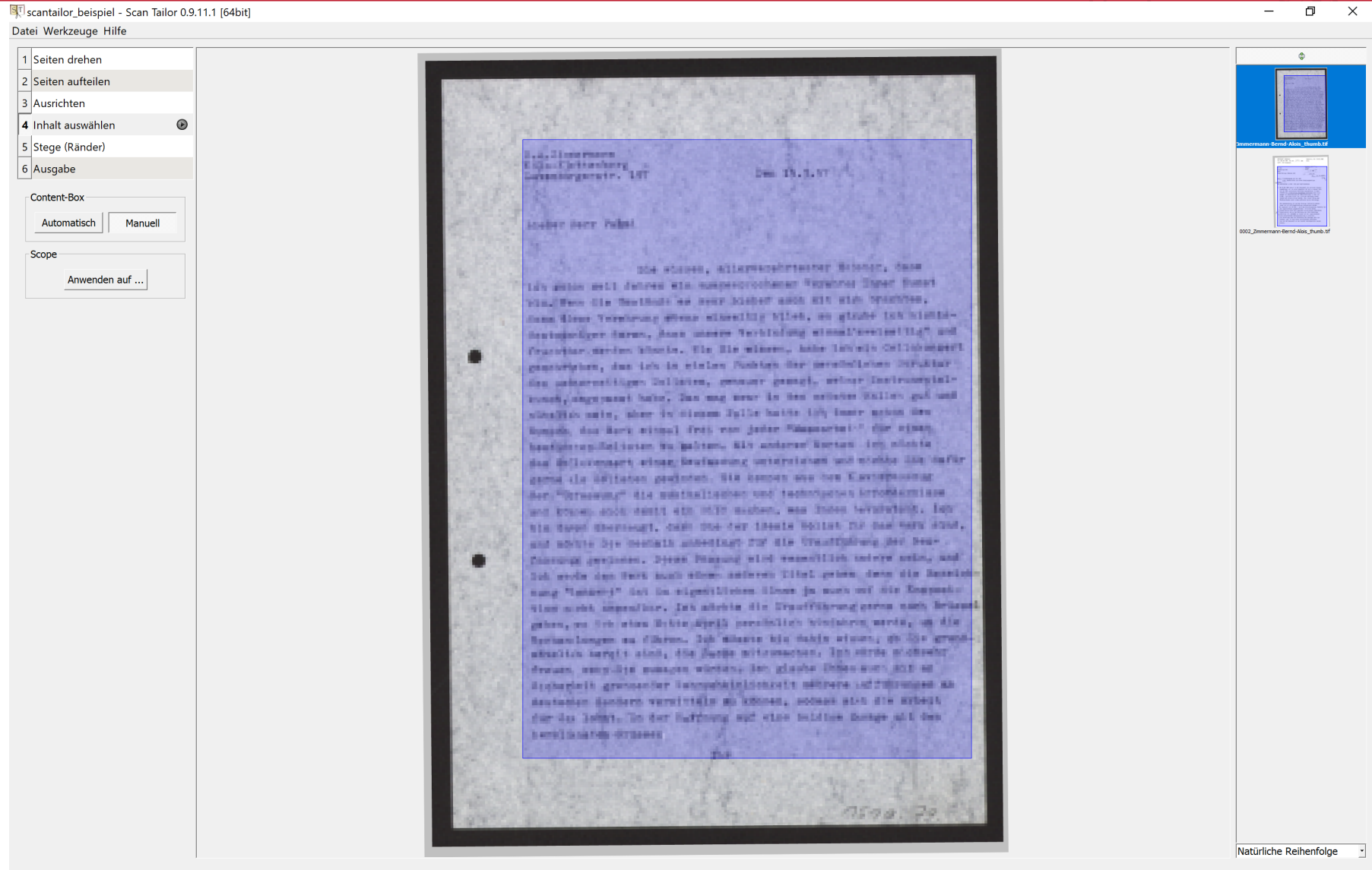


# Bildvorverarbeitung: ScanTailor

- umfassendes, frei verfügbares Werkzeug  
<https://github.com/scantailor/scantailor>
  - + graphische Benutzerschnittstelle (GUI)
  - + Kommandozeileninterface (CLI)
  - keine Programmierschnittstelle (API)
- weitgehend **automatisiert**
- erlaubt Stapelverarbeitung
- manuelle Korrektur möglich



# Bildvorverarbeitung: ScanTailor



- Prozesse zur Erkennung der Struktur auf Seiten- und Dokumentebene
  - ▶ **Page Segmentation:** Lokalisierung von zusammenhängenden Text- und Nichttextbereichen
  - ▶ **Region Classification:** Typisierung von Textbereichen
  - ▶ **Line/Character Splitting:** Lokalisierung der einzelnen Zeilen/Zeichen
  - ▶ **Document Analysis:** Konstruktion der logischen Dokumentstruktur (METS!)
- entscheidend für die korrekte **Rekonstruktion des Textflusses** und damit für maschinelle Auswertungen



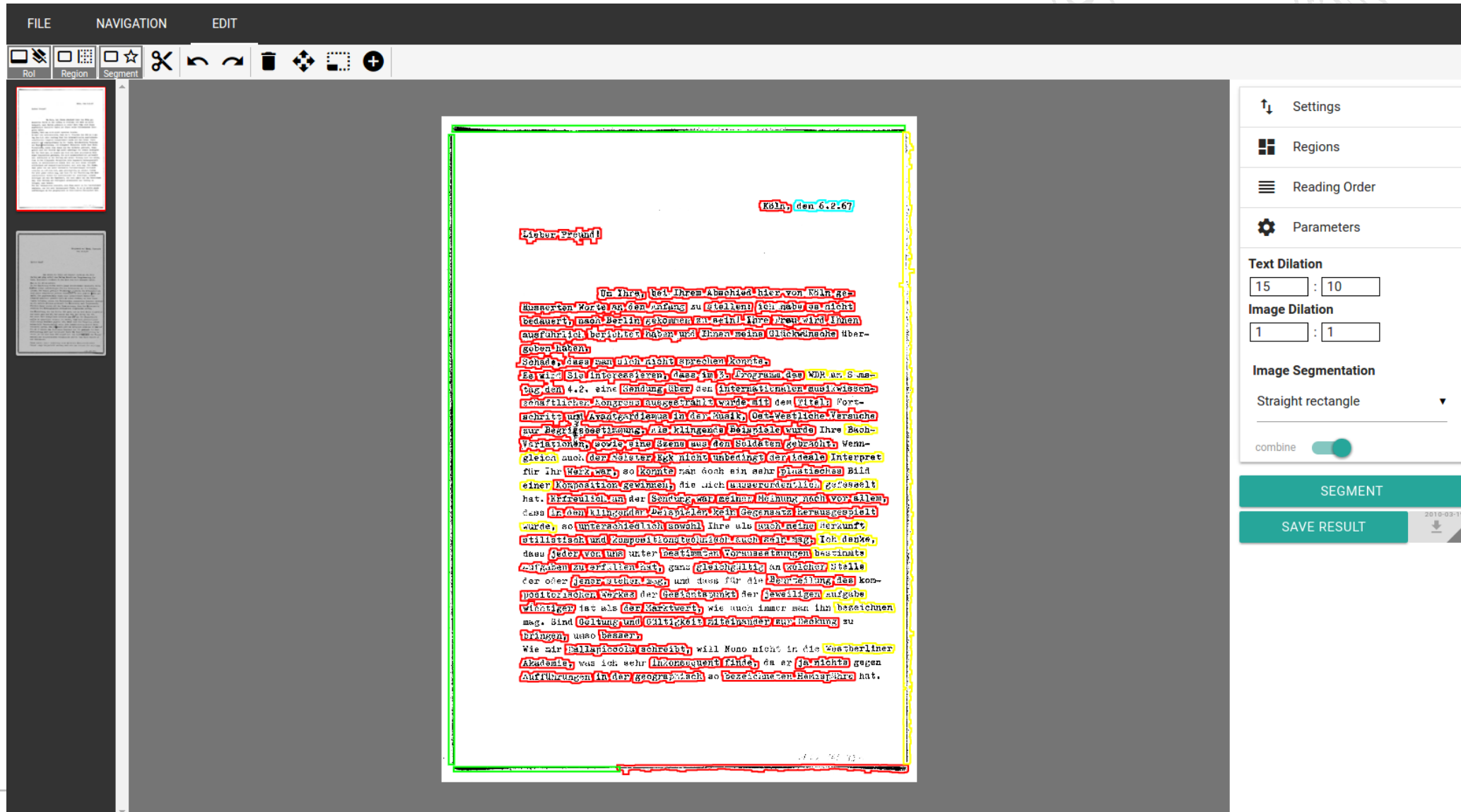
# Layoutanalyse: LAREX

- umfassendes, frei verfügbares Werkzeug  
<https://github.com/chreul/LAREX>
  - + graphische Benutzerschnittstelle (GUI)
  - Kommandozeileninterface (CLI)
  - keine Programmierschnittstelle (API)
- teilweise **automatisiert**
- erlaubt Stapelverarbeitung
- manuelle Korrektur **nötig**

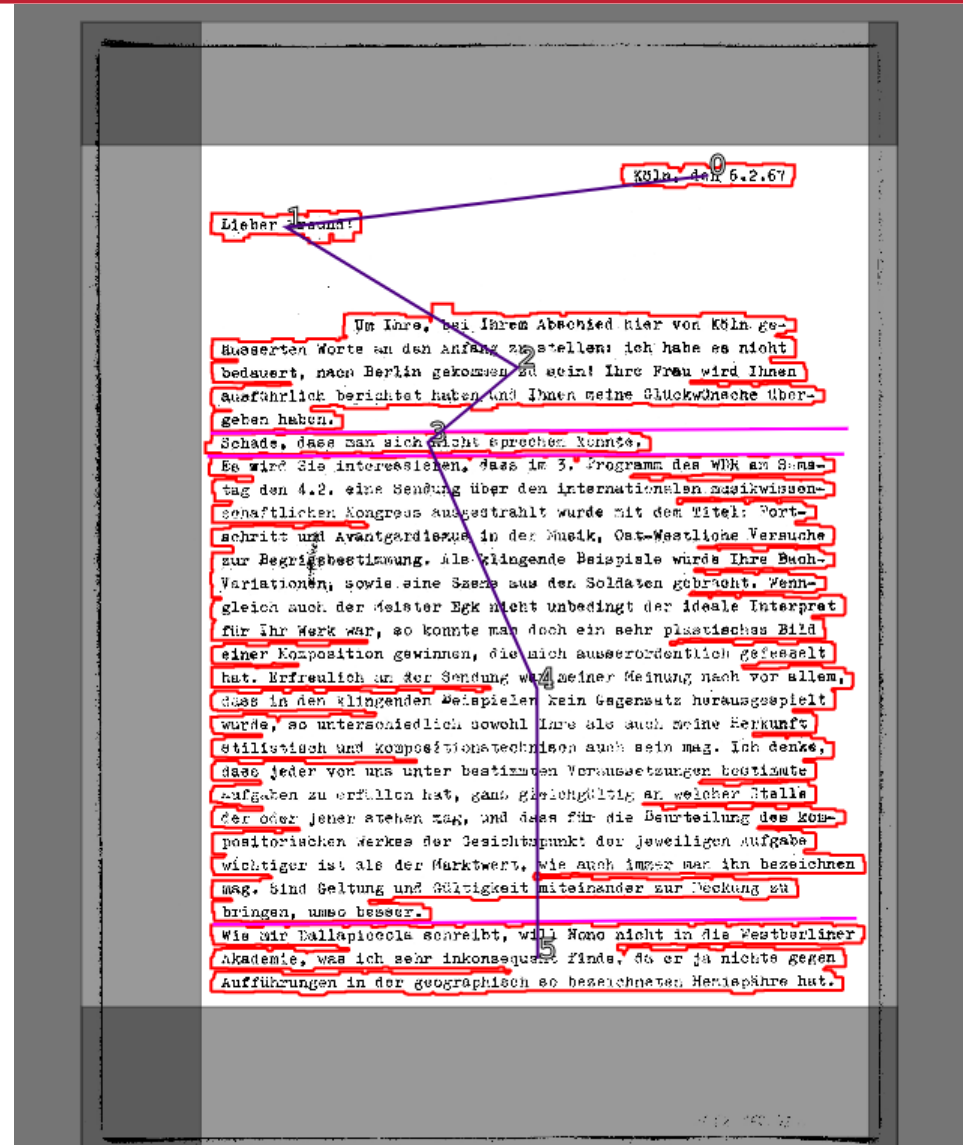




# Bildvorverarbeitung: LAREX



# Bildvorverarbeitung: LAREX



- Kernkomponente der OCR
- Genauigkeit beeinflusst vom Typ des zugrundeliegenden **Algorithmus** und vom eingesetzten **Modell**
- aktuell Paradigmenwechsel: **zeichenorientiert** → **zeilenorientiert**
  - ▶ **Deep learning:** Tiefe (i.e. vielschichtige) neuronale Netzwerke zur Sequenzklassifizierung *(Hochreiter und Schmidhuber 1997)*
  - ▶ wesentlich weniger anfällig für **Zeichenvarianz**
  - ▶ eingebautes **Sprachmodell**
- auch schwierige historische Vorlagen in „OCR-Reichweite“ *(Springmann 2016)*



# Zeilenorientierte Zeichenerkennung

- Bearbeitungsebene sind **Glyphensequenzen**

**Skalierung:** einheitliche Höhe für alle Textzeilen

**Merkmalsextraktion:** fixe Anzahl horizontaler Zeilen, variable Anzahl vertikaler Spalten:  
Textzeilen als binärwertige Vektoren

Anfangsgründe der physischen Geographie.

- kontextabhängige (i.e. *Übergangswahrscheinlichkeiten*) Erkennung (größere Mengen an Trainingsmaterial nötig als bei zeichenorientierten Verfahren)
- Zeilenerkennung als notwendiger Vorverarbeitungsschritt
- Glyphenlokalisierung geschieht en passant
- Einsatz neuronaler Netze für den Klassifizierungsschritt



# Zeichenerkennung: Textvereinigung

- Prozesse zur **Vereinigung** verschiedener OCR-Ergebnisse in einen **Volltext**
- Fehler auch bei „optimaler“ Vorverarbeitung und Verwendung spezifischer Modelle
- **unterschiedliche Engines** bzw. Modelle haben **unterschiedliche Stärken** und machen unterschiedliche Fehler
- Idee: **Extraktion** korrekt erkannter Textbestandteile **aus mehreren OCR-Durchgängen** *(Handley 1998)*
- Vorteil: Integration vorhandener OCR ebenfalls möglich
- **Reduktion** der Anzahl der falsch erkannten Zeichen um 14 % erzielt *(Boenig et al. 2016)*



# Zeichenerkennung: Werkzeuge

## ■ Abbyy FineReader

- ▶ kommerzielle Software, kaum Adaptionmöglichkeiten
- ▶ „Platzhirsch“: großflächiger Einsatz z.B. in Bibliotheken
- ▶ GUI mit Möglichkeit der Stapelverarbeitung

## ■ Tesseract

- ▶ Open Source Software mit großer Entwicklercommunity
- ▶ Integration zeilenorientierter Erkennung in Version 4
- ▶ CLI und API mit weitreichenden Adaptionmöglichkeiten

*(Smith 2007)*

## ■ OCRMerger

- ▶ Eigenentwicklung
- ▶ Textvereinigung auf Basis von `diff`
- ▶ Konfliktauflösung mit Hilfe von *Ground Truth*

*(Boenig et al. 2016)*



# Zeichenerkennung: Ergebnisse



Evaluation anhand dreier manuell transkribierter Briefe

	FineReader		Tesseract		Merged
errors	302		297		188
missing	74		0		0
total	5856		5856		5856
err	5,157 %		5,072 %		3,210 %
errnomiss	3,893 %		5,072 %		3,210 %
11	s z	19	, .	9	, .
10	¬ -	4	m n	5	—
6	o e	3	s _	4	¬ -
5	« .	3	. _	4	s z
5	* .	3	z g	3	o e

**Reduktion der  
Zeichenfehler  
um ca. 37 %.**





- kritische **Textauswahl**
- Korrektur und tiefere **Erschließung**
  - ▶ Basis: Text-Bild-Ansicht
  - ▶ Auszeichnung von Personen, Orten, Querverweisen etc.
  - ▶ Erläuterungen und Anmerkungen zu spezifischen Sachverhalten und wissenschaftlichen Fragestellungen
- Transformation in spezifisches **Editionsformat**
  - ▶ Idealfall: nachnutzbares Standardformat (z.B. DTABf)
  - ▶ Normalfall: Verlagsvorgaben

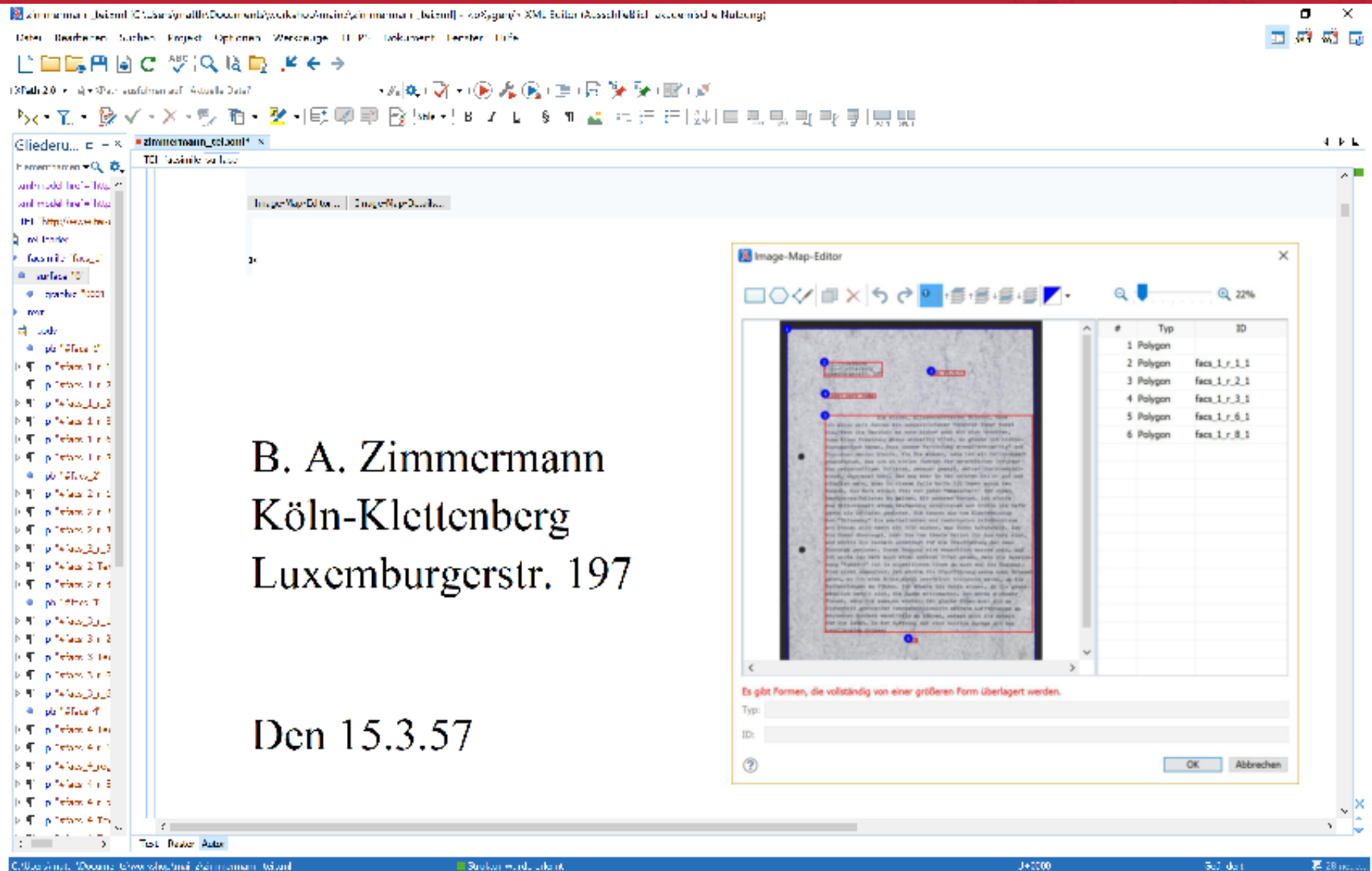


# Textbearbeitung: Oxygen

- kommerzieller XML-Editor
  - ▶ Marktführer mit großer Verbreitung (vgl. ediarum)
  - ▶ weitgehende Anpassungsmöglichkeiten
  - ▶ plattformübergreifend
- **Unterstützung** des Editionsprozesses durch Text-Bild-Ansicht unter Einbeziehung struktureller Annotationen
- **Operationalisierung** des Editionsprozesses
  - ▶ spezielle Frameworks
  - ▶ Transformationsszenarien
  - ▶ Einbindung spezieller Schematronregeln



# Textbearbeitung: Oxygen



B. A. Zimmermann  
Köln-Klettenberg  
Luxemburgerstr. 197

Den 15.3.57





# Perspektiven



- bisher kein spezifisches Modelltraining
  - ▶ „echtes“ Trainingsmaterial aus manuell erfassten Dokumenten
  - ▶ synthetisches Trainingsmaterial aus Text + Computerfont
- bisher kein spezifisches „Zimmermann-Vokabular“
  - ▶ genauere Zielhypothese bei der Textvereinigung
- bisher keine OCR-Nachkorrektur
  - ▶ PoCoTo
- bisher keine Voting-Verfahren
  - ▶ OCRopus als mögliche weitere OCR-Option

*(Vobl et al. 2014)*

*(Breuel 2008)*



# OCR-D: Überblick

- **DFG-Initiative** zur Verbesserung von OCR-Methoden für historische Drucke insbesondere für die Volltextdigitalisierung aller in den *Verzeichnissen der im deutschen Sprachraum erschienen Drucke* (VD16, VD17, VD18) nachgewiesenen Exemplare
- **Koordinationsprojekt** an der Herzog-August Bibliothek Wolfenbüttel, der Staatsbibliothek Berlin, dem Karlsruher Institut für Technologie und der BBAW → Implementierung einer Ausschreibung für methodische Projekte auf allen Ebenen eines optimierten OCR-Workflows
  - ▶ Bildvorverarbeitung
  - ▶ Layoutanalyse
  - ▶ Texterkennung/-optimierung
  - ▶ Modelltraining
  - ▶ Langzeitarchivierung
  - ▶ Qualitätssicherung



# OCR-D: Projektprämissen

- **Lückenschluss** zwischen Forschung und Praxis
  - ▶ Transfer der Forschungsergebnisse
  - ▶ zugängliche und nachnutzbare Implementierungen
- **Methodenpluralismus**
  - ▶ insbesondere bei schwierigen Vorlagen: **kein** bester Algorithmus
  - ▶ Implementierung möglichst **vieler Ansätze** samt „Auswahlmechanismus“
- konsequent **Open-Source**
  - ▶ Veröffentlichung des Quellcodes **und**
  - ▶ Anschluss an vorhandene Communities





# Editionsunterstützung

## ■ Text

- ▶ Erfassung **großer Textmengen**
- ▶ **kontinuierliche** Erfassung innerhalb des Förderungszeitraumen (on Demand)
- ▶ **geringe Kosten** nach Einrichtung des Workflows (Modelltraining, Framework)

## ■ Forschungsdaten

- ▶ Erfassung der Primärdaten **ohne Vorauswahl**
- ▶ Erschließung der Texte durch **computerlinguistische Methoden**
- ▶ schreibweisentolerante Suche
- ▶ quantitative Datenanalyse

## ■ Edition

- ▶ Auswahl spezifischer und repräsentativer Korrespondenzbestandteile
- ▶ **Aufwertung** durch die große Menge der zur Verfügung stehenden Forschungsdaten
- ▶ **Dynamisierung** durch die Möglichkeit der situativen Erweiterung





**Danke für Ihre Aufmerksamkeit!**

