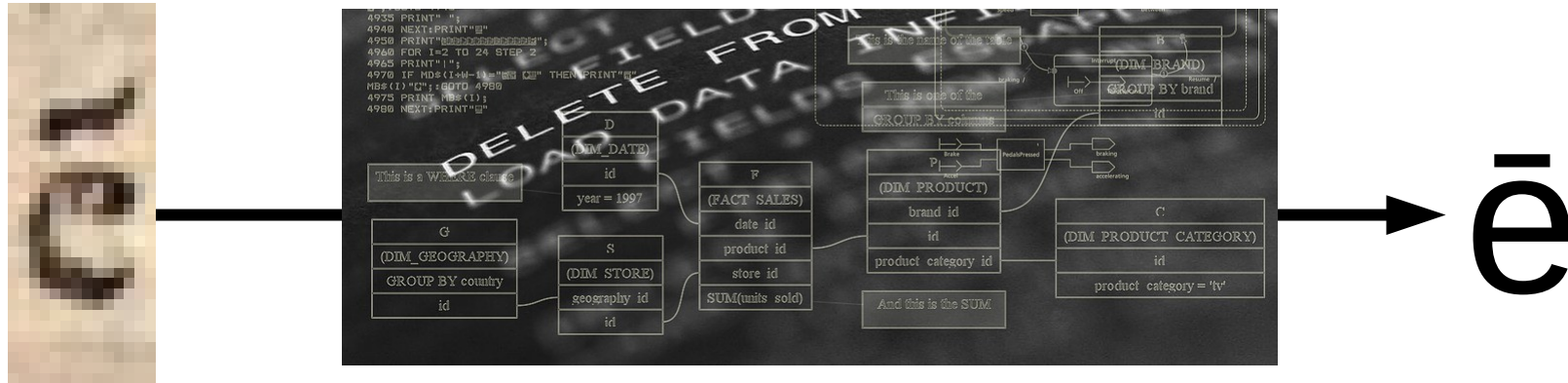


# Mit freier Software Text in Digitalisaten erkennen



## OCR-Praxis an der UB Mannheim

Stefan Weil, Philipp Zumstein

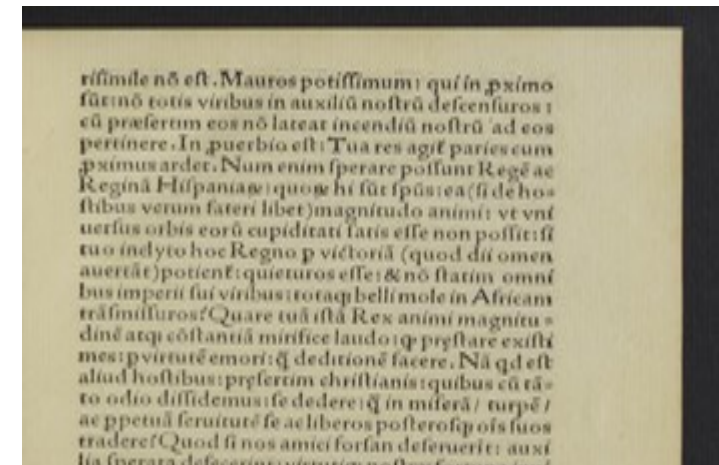
Goobi Workshop, 2016-05-12

# Übersicht

- Aktuelle **Projekte** mit OCR
- OCR-**Software**
- **Workflow**-Einbindung
- **Erkennungsgenauigkeit** messen und verbessern
- Resümee

# Aktuelle Projekte mit OCR

- Digitalisierungsprojekt "Ancien Droit"
- Projekt "Aktienführer 2"
- Reichs- und Staatsanzeiger
- ... stellen ganz unterschiedliche Anforderungen an die OCR



(Umsatz Mittel)	5 119	5 117	Abschreibungen	5 119	5 117
(Eigenkapital)	5 117	5 117	BV-Hausen	500	500
(Fremdkapital)	5 118	5 118	Jahresüberschuss	241	188

### BASF Aktiengesellschaft

**Adresse:** 6700 Ludwigshafen (Rhein)  
**Telefon:** (09 31) 5 31  
**Telefax:** 444 811

**Vorstand:**  
 Prof. Dr. rer. nat. Matthias Goetschalckx, Ludwigshafen (Rhein), Vorsitz  
 Dr. rer. nat. Hans Mühl, Ludwigshafen (Rhein), stellv. Vorsitz  
 Dr. rer. nat. Hans Ahrens, Ludwigshafen (Rhein)  
 Dr. rer. nat. Bernd Harnack, Ludwigshafen (Rhein)  
 Dr. Ing. Reich Henkel, Ludwigshafen (Rhein)  
 Dr. rer. nat. Wolfgang Jantschek, Ludwigshafen (Rhein)  
 Prof. Dr. Ing. Horst Pannier, Ludwigshafen (Rhein)  
 Dr. Ing. Hans August Wegjen, Ludwigshafen (Rhein)  
 Dr. rer. nat. Herbert Wilberstein, Ludwigshafen (Rhein)  
 Hans Joachim Witt, Ludwigshafen (Rhein)

**1. Aufsichtsratsvorsitzender:**  
 Dr. jur. Robert Biers, Königstein/Taunus  
 Dr. rer. nat. Manfred Rigen, Ludwigshafen  
 Prof. Dr. Ing. Gerhard Frank, Heilbronn  
 Dr. rer. nat. Johan M. Goutschaert, Wassenaar (Niederlande)  
 Dr. rer. Wolfgang Heeseler, Heilbronn  
 Karl Herrmann, Cuxhaven/Pöhl (1)  
 Dr. rer. nat. Hans-Joachim Langemann, Frankfurt  
 Dr. jur. Robert Holsch, Jena (1)  
 (1) Dr. rer. nat. Hans-Joachim Langemann, Jena  
 (2) Dr. rer. nat. Hans-Joachim Langemann, Jena  
 (3) Dr. rer. nat. Hans-Joachim Langemann, Jena  
 (4) Dr. rer. nat. Hans-Joachim Langemann, Jena  
 (5) Dr. rer. nat. Hans-Joachim Langemann, Jena  
 (6) Dr. rer. nat. Hans-Joachim Langemann, Jena  
 (7) Dr. rer. nat. Hans-Joachim Langemann, Jena  
 (8) Dr. rer. nat. Hans-Joachim Langemann, Jena  
 (9) Dr. rer. nat. Hans-Joachim Langemann, Jena  
 (10) Dr. rer. nat. Hans-Joachim Langemann, Jena

**2. Aufsichtsratsmitglieder:**  
 Dr. rer. nat. Hans-Joachim Langemann, Jena  
 Dr. rer. nat. Hans-Joachim Langemann, Jena  
 Dr. rer. nat. Hans-Joachim Langemann, Jena  
 Dr. rer. nat. Hans-Joachim Langemann, Jena  
 Dr. rer. nat. Hans-Joachim Langemann, Jena  
 Dr. rer. nat. Hans-Joachim Langemann, Jena  
 Dr. rer. nat. Hans-Joachim Langemann, Jena  
 Dr. rer. nat. Hans-Joachim Langemann, Jena  
 Dr. rer. nat. Hans-Joachim Langemann, Jena  
 Dr. rer. nat. Hans-Joachim Langemann, Jena  
 Dr. rer. nat. Hans-Joachim Langemann, Jena

**Gründungsdatum:** 8. 8. 1865, l. G. Fortsetzung: 21. 1. 1926; Fortsetzung im Jahr der l. G. Fortsetzung 20. 1. 1926.



# Projekt "Ancien Droit"

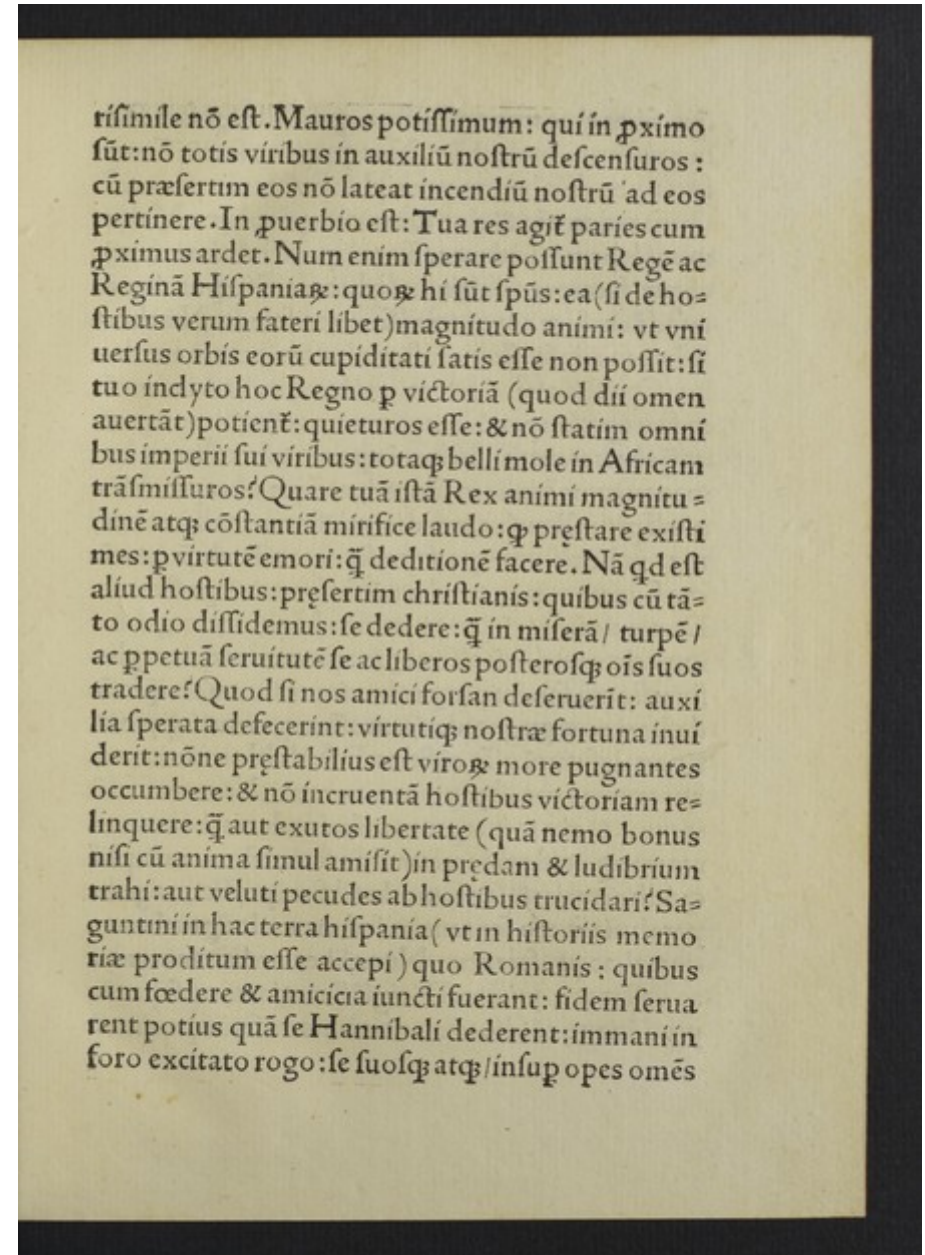
800 rechtshistorische Titel des 16. bis 18. Jhd. aus der Sammlung Desbillons werden digitalisiert und mit computerlinguistischen Verfahren analysiert.

## Besonderheiten

- Antiqua-Schriften, aber mit speziellen Zeichen wie langes S, Ligaturen u. a.
- Hauptsächlich Alt-Französisch und Latein

## Anforderungen

- Hohe Erkennungsgenauigkeit, insbesondere Wortgenauigkeit



# Projekt "Aktienführer 2"

Saling's Börsen-Jahrbuch und Hoppenstedt Aktienführer von 1880 bis 1978

## Besonderheiten

- Antiqua-Schrift
- Deutsch, aber mit internationalen Firmen- und Personennamen

## Anforderungen

- Hohe Erkennungsgenauigkeit, insbesondere bei Zahlen
- Layouterkennung (Tabellen) besonders wichtig

	1976	1977
<b>Dividenden auf Stammaktien:</b>		
1973: 8 % (Div. Sch. Nr. 24)		
1974 u. 1975: je 4 % (Div. Sch. Nr. 25, 26)		
1976: 8 % (Div. Sch. Nr. 27)		
1977: 6, - DM + 3,38 DM St.G. (Div. Sch. Nr. 28)		
<b>Aus den Bilanzen (in 1 000 DM)</b>		
	1976	1977
Anlagevermögen	4 395	5 305
Umlaufvermögen (flüssige Mittel)	10 260	10 788
Eigenkapital (Grundkapital)	5 117	5 117
	3 015	3 015
<b>B</b>		
	1976	1977
Fremdkapital	9 360	10 833
Bilanzgewinn	242	185
Bilanzsumme	14 655	16 094
<b>Aus den Gewinn- und Verlustrechnungen</b>		
	1976	1977
Umsatzerlöse	17 910	22 106
Materialaufwand	7 510	8 579
Personalaufwand	9 237	10 389
Abschreibungen	810	767
EEV-Steuern	299	539
Jahresüberschuß	241	184

## BASF Aktiengesellschaft



Sitz: 6700 Ludwigshafen (Rhein)  
Telefon: (06 21) 6 01  
Telex: 4 64 811

### Vorstand:

Prof. Dr. rer. nat. Matthias Seefelder, Ludwigshafen (Rhein), Vors. ;  
Dr. rer. nat. Hans Moell, Ludwigshafen (Rhein), stellv. Vors. ;  
Dr. rer. nat. Hans Albers, Ludwigshafen (Rhein);  
Dr. rer. pol. Ernst Denzel, Ludwigshafen (Rhein);  
Dr. -Ing. Erich Henkel, Ludwigshafen (Rhein);  
Dr. rer. nat. Wolfgang Jentsch, Ludwigshafen (Rhein);  
Prof. Dr. -Ing. Horst Pommer, Ludwigshafen (Rhein);  
Dr. -Ing. Karl August Wetjen, Ludwigshafen (Rhein);  
Dr. rer. nat. Herbert Willersinn, Ludwigshafen (Rhein);  
Hans Joachim Witt, Ludwigshafen (Rhein)

### Aufsichtsrat:

Prof. Dr. phil. nat. Bernhard Timm, Heidelberg, Vors. ;  
Werner Vitt, Isernhagen, stellv. Vors. +);  
Dr. rer. nat. Wolfgang Arend, Ludwigshafen +);

Dr. jur. Robert Ehret, Königstein/Taunus;  
Prof. Dr. rer. nat. Manfred Eigen, Göttingen;  
Prof. Dr. -Ing. Berthold Frank, Heidelberg;  
Dr. rer. pol. Johan M. Goudswaard, Wassenaar (Niederlande);  
Dr. jur. Wolfgang Heintzeler, Heidelberg;  
Kurt Herrmann, Carlsberg/Pfalz +);  
Dr. rer. pol. Kurt Hohenemser, Frankfurt;  
Dr. jur. Robert Holzach, Zumikon (Schweiz);  
Christoph von Knorre, Ludwigshafen +);  
Roland Koch, Ludwigshafen +);  
Herbert Krug, Ludwigshafen +);  
Dr. rer. nat. Hans Joachim Langmann, Jugenheim/Bergstraße;  
Prof. Dr. phil. h. c. Hans L. Merkle, Stuttgart;  
Heinz-Werner Meyer, Dortmund +);  
Wilhelm Roßmüller, Marl +);  
Willi Schüler, Heringen +);  
Rudolf Woll, Mainz +)

+ ) Arbeitnehmervertreter

Gründung: 6. 4. 1865, I. G. Farbenindustrie AG seit 1925; Neugründung im Zuge der I. G. -Entflechtung 30. 1. 1952.

### Tätigkeitsgebiet:

Die Interessen der BASF reichen von den chemischen Rohstoffen über alle Veredelungsstufen bis zu den Endprodukten und ihren Anwendungen. Das entsprechend vielseitige

109

# Reichs- und Staatsanzeiger

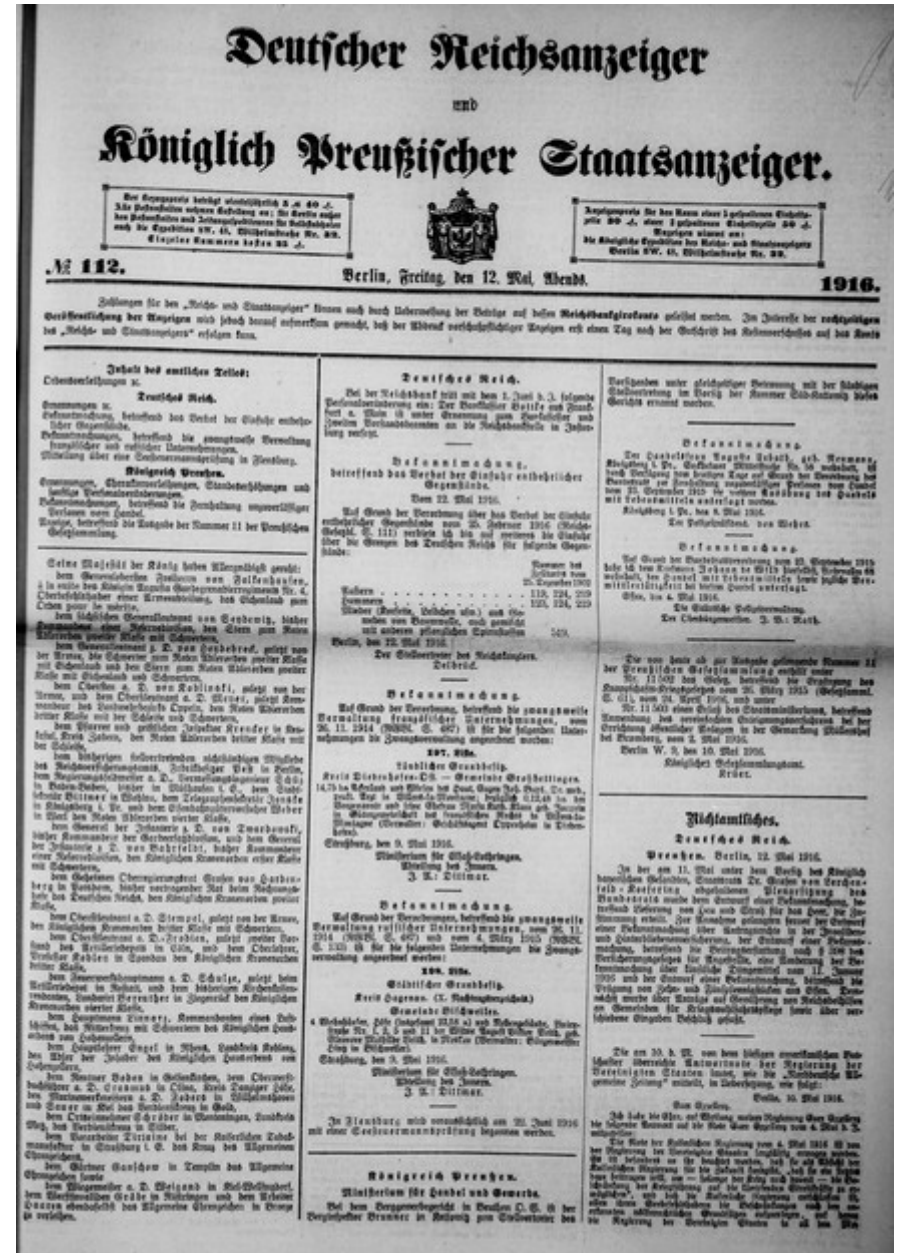
(Fast) alle Ausgaben von 1819  
(Allgemeine Preußische Staats-  
Zeitung) bis 1945 (Deutscher  
Reichsanzeiger und Preußischer  
Staatsanzeiger).

## Besonderheiten

- Fraktur-Schrift
- Scans von Mikrofilmen in teilweise mäßiger Qualität
- Menge (127 Jahre, über 38000 Ausgaben, 25 TB TIFF Scans)

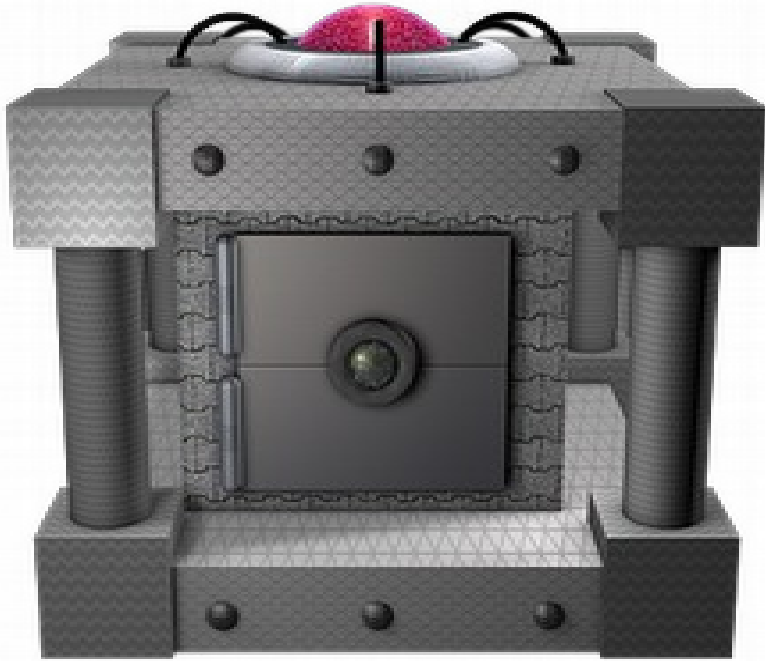
## Ziel

- Erschließung für (unscharfe) Suche nach Stichworten



# OCR Software

## Kommerzielle Software



ABBYY Finereader  
BIT Alpha

## Freie Software



Tesseract  
Ocropus  
CuneiForm  
Ocrad

# ABBYY Fine Reader

- Kommerzielle Software für Windows und Linux
- ABBYY OCR SDK, Cloud OCR SDK oder Linux CLI
- Beispiel: FineReader Engine 11 CLI for Linux, 120'000 Seiten / Jahr, 999 EUR einmalig
- Unterstützung für Fraktur (OldGerman, OldFrench usw., kein OldLatin!) erfordert (teure) Projektlizenz
- Ausgabeformate TXT, ALTO, XML, PDF, u. a.
- Zeichenerkennung + Wörterbuch (beides sprachabhängig) mit sehr starker Gewichtung des Wörterbuchs
- Training mit Windows-Version und OCR unter Linux?
- <http://www.ocr4linux.com/>



# Tesseract

- Grafikdatenformate TIFF, PNG, JPEG, JPEG2000, ...
- Layouterkennung mit [Leptonica](#)
- Zeichenbasierte Erkennung durch Mustervergleich
- Geplant für 2016: neuronales Netzwerk
- Mehr als 100 Sprachen auswählbar
- Sprachregeln (Wörterbücher, Silben, ...) werden nur als Hinweise verwendet
- Ausgabeformate hOCR, TXT, PDF und Spezialformate
- Sehr aktive Entwickler-Community  
<https://github.com/tesseract-ocr>
- Bestandteil aller großen Linux-Distributionen
- Freie Nutzung z. B. für Distributed OCR ([Bachelor-Arbeit](#))

# OCRopus / OCRopy

- “Baukasten”-Philosophie: viele kleine Tools für Teilaufgaben
- Zeichenerkennung durch neuronales Netz
- Kein Wörterbuch
- Training sehr wichtig
- Modelle für Antiquaschriften und Fraktur
- Ausgabeformat hOCR
- <https://github.com/tmbdev/ocropy>

# OCR / Volltext im DFG-Viewer

The screenshot displays the DFG-Viewer interface. At the top left, the logo 'DFGviewer' is visible. In the top right corner, there is a link 'Mehr zum DFG-Viewer: DFG-Viewer' and the DFG logo. Below the header, the document title is '[Brief von Christian von Ehrenfels an Otto Selz vom 06. April 1914] Prag 1914' and the URL 'urn:nbn:de:bsz:180-digosi-616' is shown. The main content area is split into two parts: a scanned image of the document on the left and a text box on the right containing the OCR result. The scanned document shows a handwritten '3.3.' in the top right, a date 'Prag, den 6. April 1914.' in the center, and a salutation 'Sehr geehrter Herr Kollege !' in purple. The main body of the text is highlighted in orange and matches the OCR text on the right. The OCR text is in a serif font and contains several underlined words: 'Psychologie' and 'Experimentalpsychologie'. The interface includes navigation icons and a 'Seite 10' indicator at the bottom.

DFGviewer

Mehr zum DFG-Viewer: DFG-Viewer |

Titel: [Brief von Christian von Ehrenfels an Otto Selz vom 06. April 1914] Prag 1914  
URN: urn:nbn:de:bsz:180-digosi-616

Seite 10

Volltext markieren

3.3.

Prag, den 6. April 1914.

Sehr geehrter Herr Kollege !

Der Schreiber dieser Zeilen ist Mitglied der Kommission, welche zur Erstattung eines Vorschlages für die Wiederbesetzung der an der deutschen Universität in Prag vakanten Professur für Philosophie gewählt wurde .

Meiner Ansicht nach (die jedoch nicht alle Kommissionsmitglieder teilen) brauchen wir vor allem einen Forscher und Lehrer, der die Psychologie in ihrer vollen Ausdehnung zu seinem Gebiet erkoren hat und auch speziell Experimentalpsychologie betreibt, - einen jungen, tatkräftigen Mann, der

Der Schreiber dieser Zeilen ist Mitglied der Kommission, welche zur Erstattung eines Vorschlages für die Wiederbesetzung der an der deutschen Universität in Prag vakanten Professur für Philosophie gewählt wurde .

Meiner Ansicht nach (die jedoch nicht alle Kommissionsmitglieder teilen) brauchen wir vor allem einen Forscher und Lehrer, der die Psychologie in ihrer vollen Ausdehnung zu seinem Gebiet erkoren hat und auch speziell Experimentalpsychologie betreibt, - einen jungen, tatkräftigen Mann, der

Seite 10

# Workflow-Einbindung

Bisher wird die OCR erst nach Veröffentlichung eines Digitalisats durchgeführt (außerhalb des Goobi-Workflows):

1) ALTO-Datei (XML mit OCR-Ergebnis) erzeugen:

```
$ abbyocr11 -rl German -if max/275308_0089.jpg \  
-f PDF -of pdf/275308_0089.pdf \  
-f ALTO -of alto/275308_0089.xml \  
-f XML -of abby/275308_0089.xml
```

oder

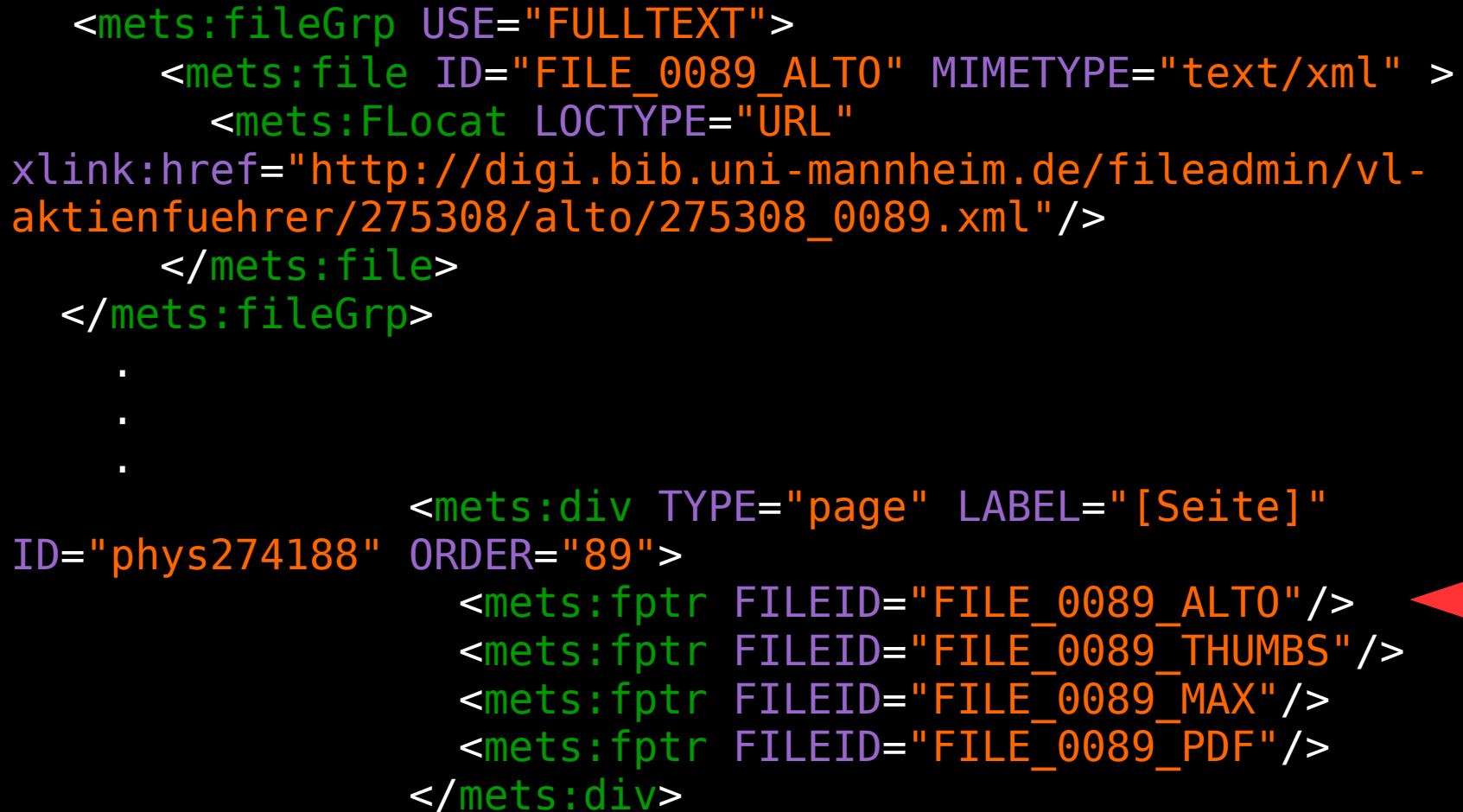
```
$ tesseract max/275308_0089.jpg hocr/275308_0089 \  
-l deu hocr  
$ ocr-transform hocr alto2.0 \  
hocr/275308_0089.hocr alto/275308_0089.xml
```

<https://github.com/UB-Mannheim/ocr-transform>

# Workflow-Einbindung

2) Verweise auf Volltext in METS/MODS-Datei ergänzen:

```
<mets:fileGrp USE="FULLTEXT">
  <mets:file ID="FILE_0089_ALTO" MIMETYPE="text/xml" >
    <mets:FLocat LOCTYPE="URL"
xlink:href="http://digi.bib.uni-mannheim.de/fileadmin/vl-
aktienfuehrer/275308/alto/275308_0089.xml"/>
    </mets:file>
  </mets:fileGrp>
  .
  .
  .
  <mets:div TYPE="page" LABEL="[Seite]"
ID="phys274188" ORDER="89">
    <mets:fptr FILEID="FILE_0089_ALTO"/>
    <mets:fptr FILEID="FILE_0089_THUMBS"/>
    <mets:fptr FILEID="FILE_0089_MAX"/>
    <mets:fptr FILEID="FILE_0089_PDF"/>
  </mets:div>
```



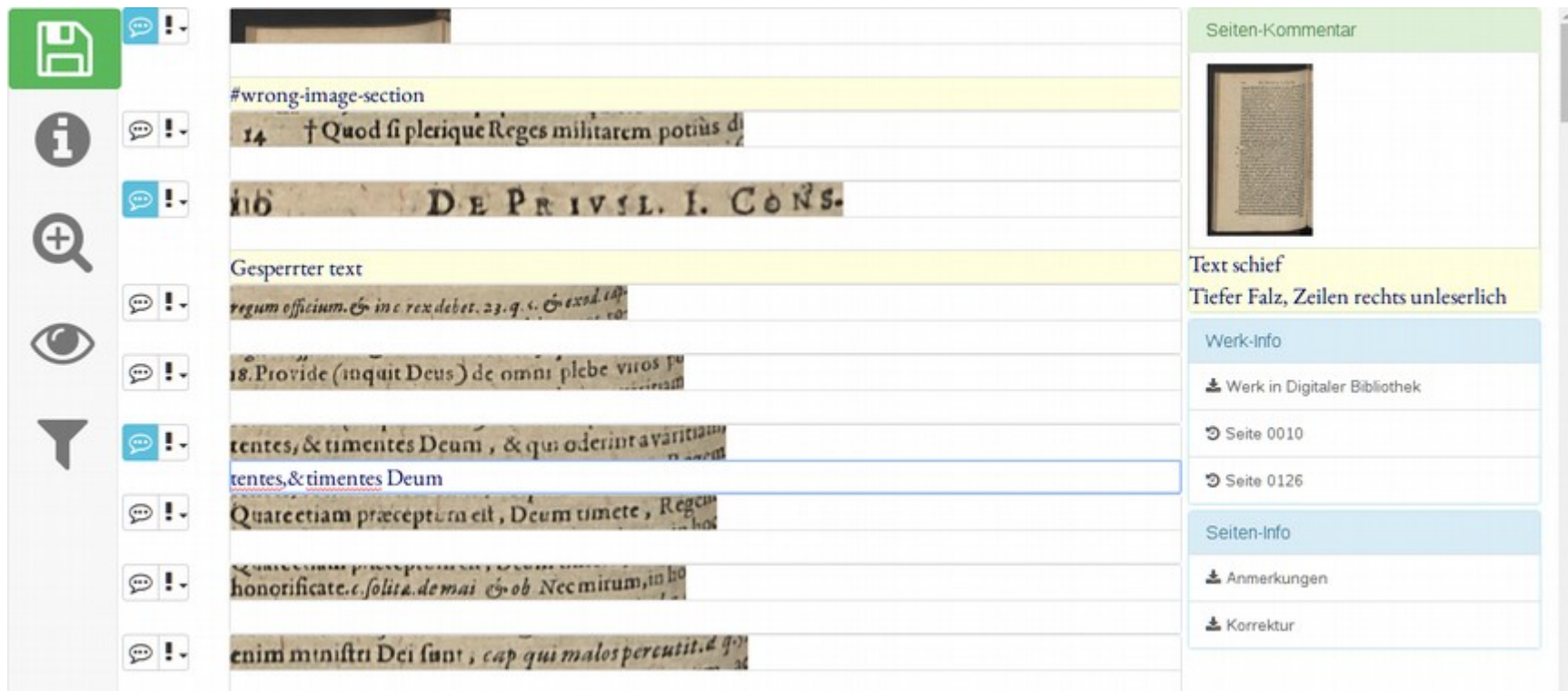
# Erkennungsgenauigkeit messen und verbessern

# Erkennungsgenauigkeit: Anwendungsfälle

- Zielvorgaben überprüfen
- OCR-Qualität eines Dienstleisters systematisch prüfen
- Entscheidungsgrundlage für weitere Optimierungsschritte
- OCR-Software optimal auf die Vorlage konfigurieren
- Trainingsdaten nebenbei erzeugen

# Erkennungsgenauigkeit messen

- Ground Truth, Gold Standard
- Ein paar Zeilen/Seiten durch BearbeiterIn erfassen lassen
  - Basis: [ocropus-gtedit](#) + [hocr-tools](#)
  - Webinterface: <https://github.com/UB-Mannheim/ocr-gt-tools>



The screenshot displays the ocr-gt-tools web interface. On the left is a vertical toolbar with icons for save, info, zoom, view, and filter. The main area shows a manuscript page with several lines of text. Some lines are highlighted in yellow, indicating errors or specific annotations. The text includes: "#wrong-image-section", "14 † Quod si plerique Reges militarem potius di", "116 DE PRIVILEGIIS. I. CONS.", "Gesperrter text", "regum officium. & in c. rex debet. 23. q. 6. & exod. cap", "18. Provide (inquit Deus) de omni plebe viros pu", "tentes, & timentes Deum, & qui oderint avaritiam", "tentes, & timentes Deum", "Quare etiam præcepta sunt, Deum timete, Regem", "honorificate. c. soluta. de mai. & ob. Nec mirum, in ho", "enim ministri Dei sunt, cap qui malos percutit. d. q. 10". On the right side, there is a sidebar with a "Seiten-Kommentar" section containing a thumbnail of the manuscript page. Below that, a "Text schief" section highlights the text "Tiefer Falz, Zeilen rechts unleserlich". Further down, a "Werk-Info" section lists "Werk in Digitaler Bibliothek", "Seite 0010", and "Seite 0126". At the bottom of the sidebar, a "Seiten-Info" section lists "Anmerkungen" and "Korrektur".



# Erkennungsgenauigkeit messen

- Vergleich mit Ergebnis der OCR
  - Zeilenweise Vergleich mit Edit-Distanz
  - `ocropus-errs` + `ocropus-econf`

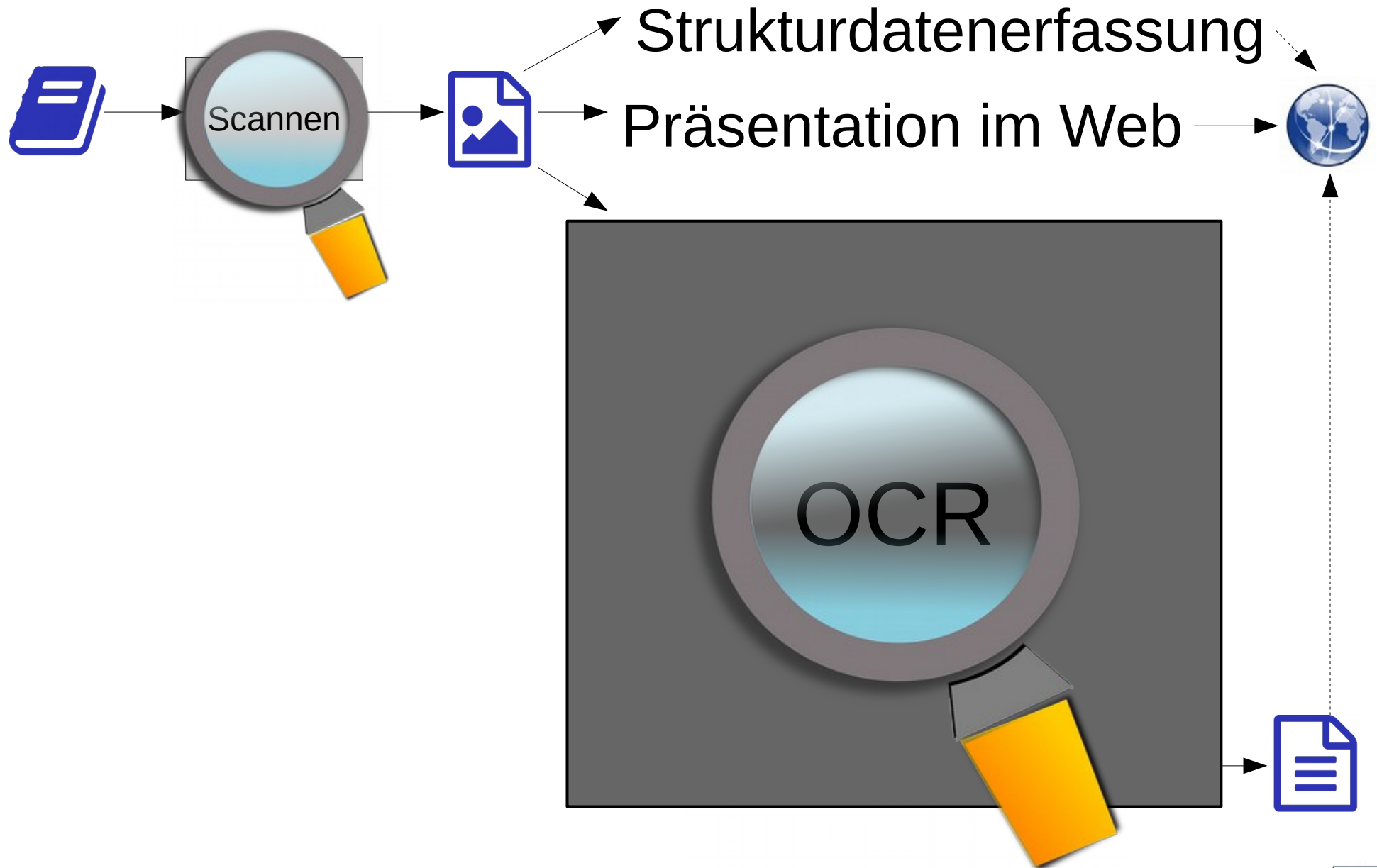
```
$ ocropus-errs *.gt.txt
```

```
errors          14
missing         0
total          2555
err             0.548 %
Errnomiss      0.548 %
```

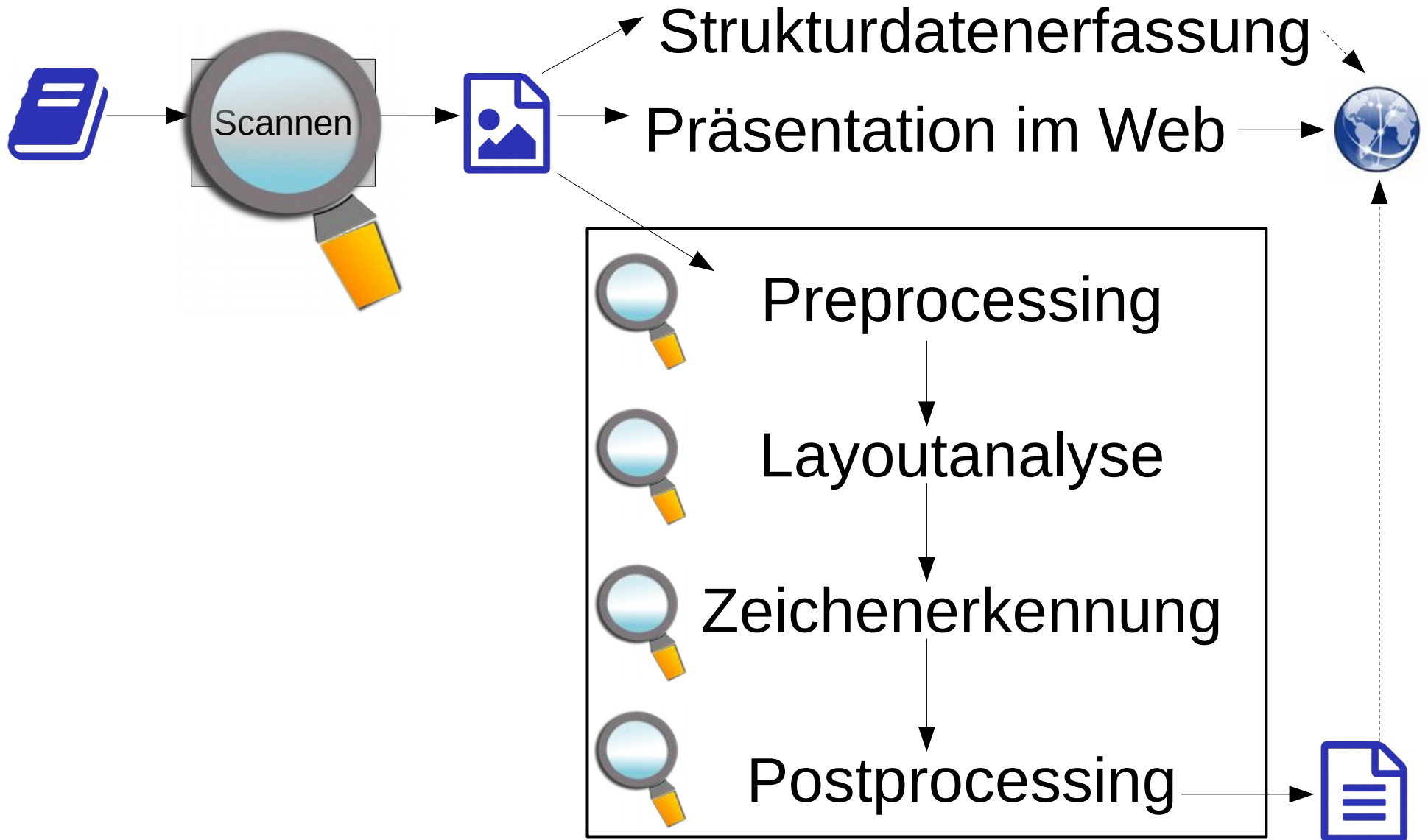
```
$ ocropus-econf *.gt.txt
```

```
1      -
1  W_  w
1  H_  fi
1  !_  -
1      -      that
1  H_  fl
1  e_  -
1  -_  -
```

# Erkennungsgenauigkeit verbessern



# Erkennungsgenauigkeit verbessern

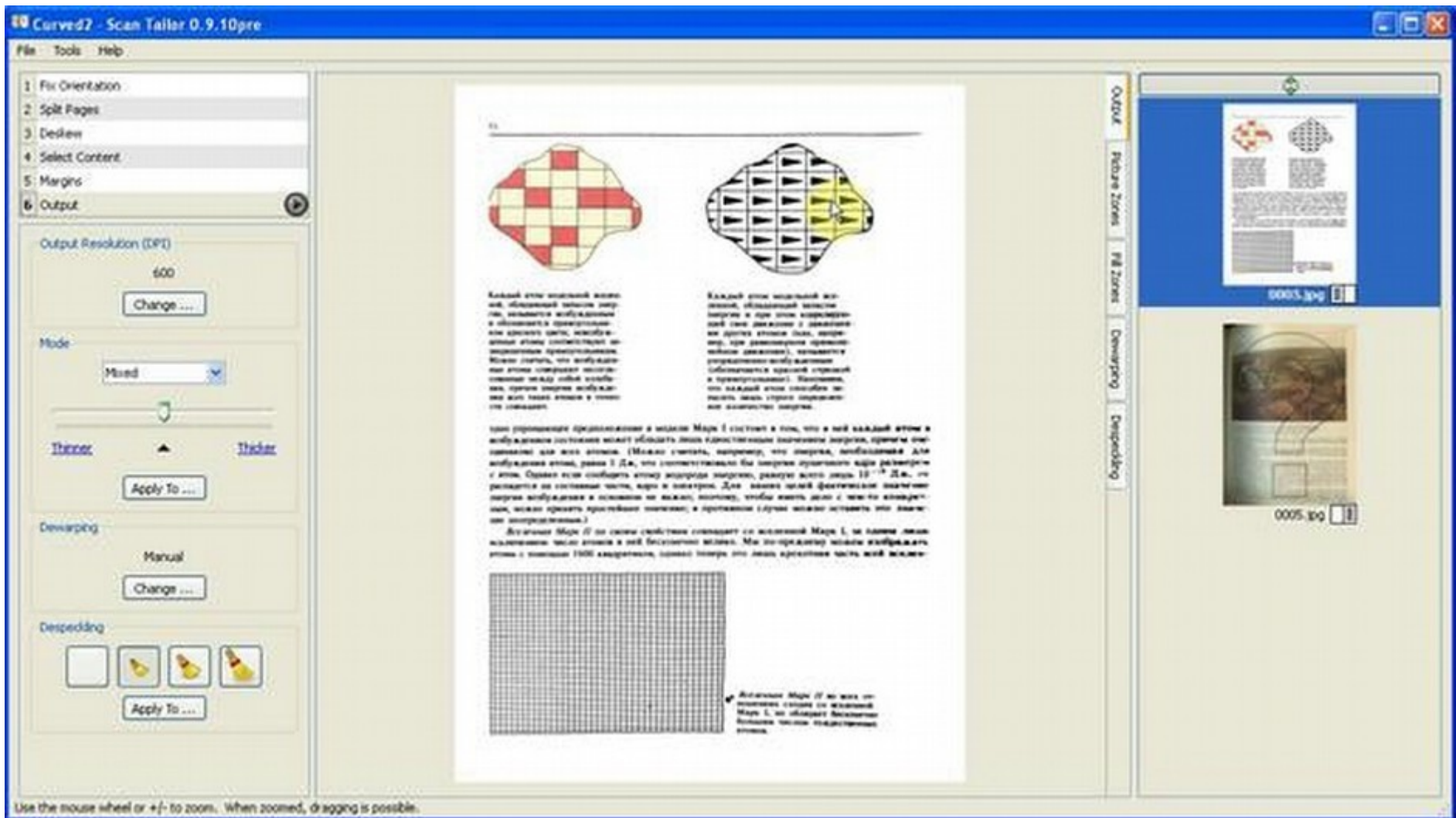


# Preprocessing

- Seitenaufteilung
- Ausrichten
- Seitenränder anpassen
- Inhalt wählen, Aufteilung Text/Bild
- Entzerren (dewarping)
- Artefakte entfernen (despeckling)
- Binarisierung

# Preprocessing: Beispiel ScanTailor

<https://github.com/scantailor/scantailor>, GPL v2, C++



# Postprocessing

- Fehler korrigieren durch Wörterbuchabgleich
- Korrektur automatisch oder semi-automatisch
- Beispiel: **PoCoTo** - Post Correction Tool (CIS München)

The screenshot shows the PoCoTo software interface. At the top, there is a window title bar with 'Page 9 of 100' and 'Causam 1/1'. Below the title bar are several buttons: 'Select page', 'Select all', 'Change candidate', and 'Correct'. A status bar below the buttons displays: 'Entries Total: 6', 'Entries Selected: 0', 'Entries Corrected: 6', 'Entries Disabled: 0', and 'Entries to process: 0'. The main area of the window contains a list of Latin text entries, each with its original form and a corrected form. The corrected forms are shown in a smaller font and are underlined. Each entry also has a 'Causam' label and a '-> Causam' button. The entries are as follows:

Original Text	Corrected Text
incedere sine vulnere .	incedere sine vulnere .
Causam	Causam
tamen cur pars	tamen cur pars
Causam	Causam
cogitationem hanc veniatur ,	cogitationem hanc veniatur ,
Causam	Causam
aliquam esse æter-	aliquam esse æter-
Causam	Causam
statim, eandem rem	statim, eandem rem
Causam	Causam
habuisse, quæ effecit,	habuisse, quæ effecit,
Causam	Causam
quem viderit, ad	quem viderit, ad
Causam	Causam
eius proximam ratiocinaretur, &	eius proximam ratiocinaretur, &
Causam	Causam
ad illius Causæ	ad illius Causæ
Causam	Causam
proximam procederet, &	proximam procederet, &
Causam	Causam
qui quid fit	qui quid fit
Causam	Causam
esse nesciunt (id	esse nesciunt (id
Causam	Causam

# Resümee

- OCR: Keine One-Size-Fits-All-Lösungen
- Freie OCR Software sind konkurrenzfähig zu kommerzieller Software, bedürfen aber mehr Konfiguration/Anpassungen
- Erkennungsgenauigkeit messen und verbessern
- Gute OCR steht und fällt mit
  - Scan-Qualität
  - Aufwand beim Pre-Processing
  - Training der OCR-Software
  - domänenspezifischem Post-Processing



Links zu OCR: <https://github.com/kba/awesome-ocr>

*Vielen Dank für die Aufmerksamkeit! Fragen? Diskussion*

# Bildquellen

- <https://pixabay.com/de/programmieren-computersprache-942487/> (Pixabay, CC0)
- <https://pixabay.com/de/sicher-metall-metallischen-ger%C3%A4t-298244/> (Pixabay, CC0)
- <https://pixabay.com/de/suchen-pr%C3%BCfen-suche-erkennen-lupen-148095/> (Pixabay, CC0)
- fa-file-text-o + fa-file-image-o + fa-book. Font Awsome (SIL OFL 1.1)
- <https://commons.wikimedia.org/wiki/File:Applications-internet.svg> (CC0)