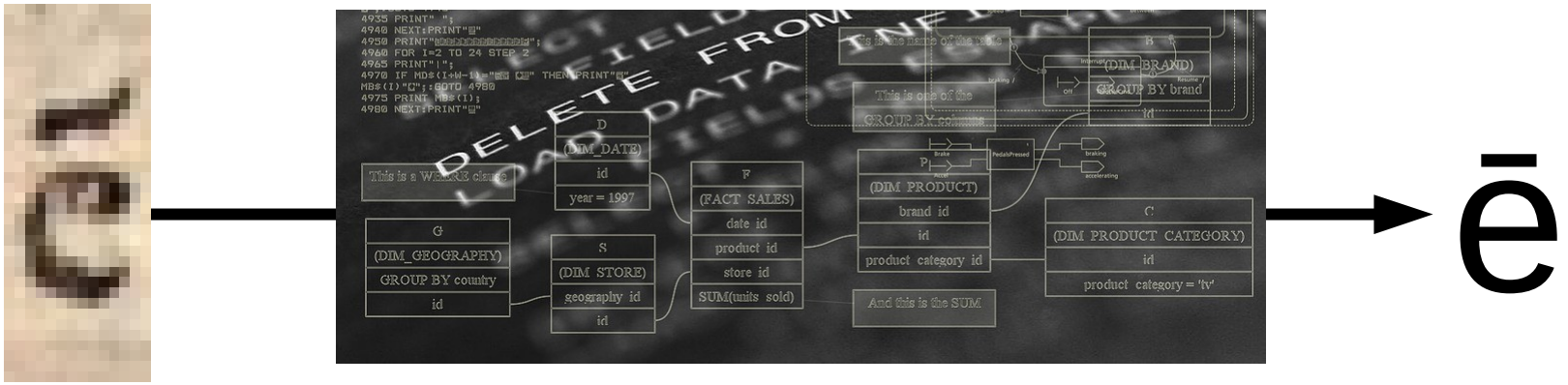
Mit freier Software Text in Digitalisaten erkennen



OCR-Praxis an der UB Mannheim

Stefan Weil, Philipp Zumstein

Goobi Workshop, 2016-05-12

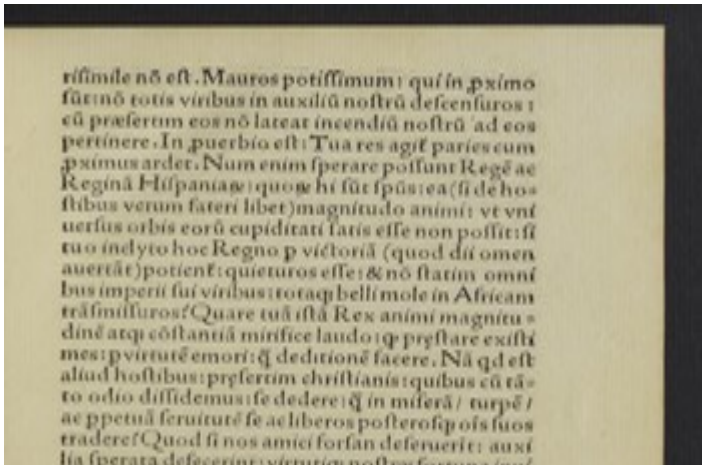
Übersicht

- 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



Projekt “Ancien Droit”

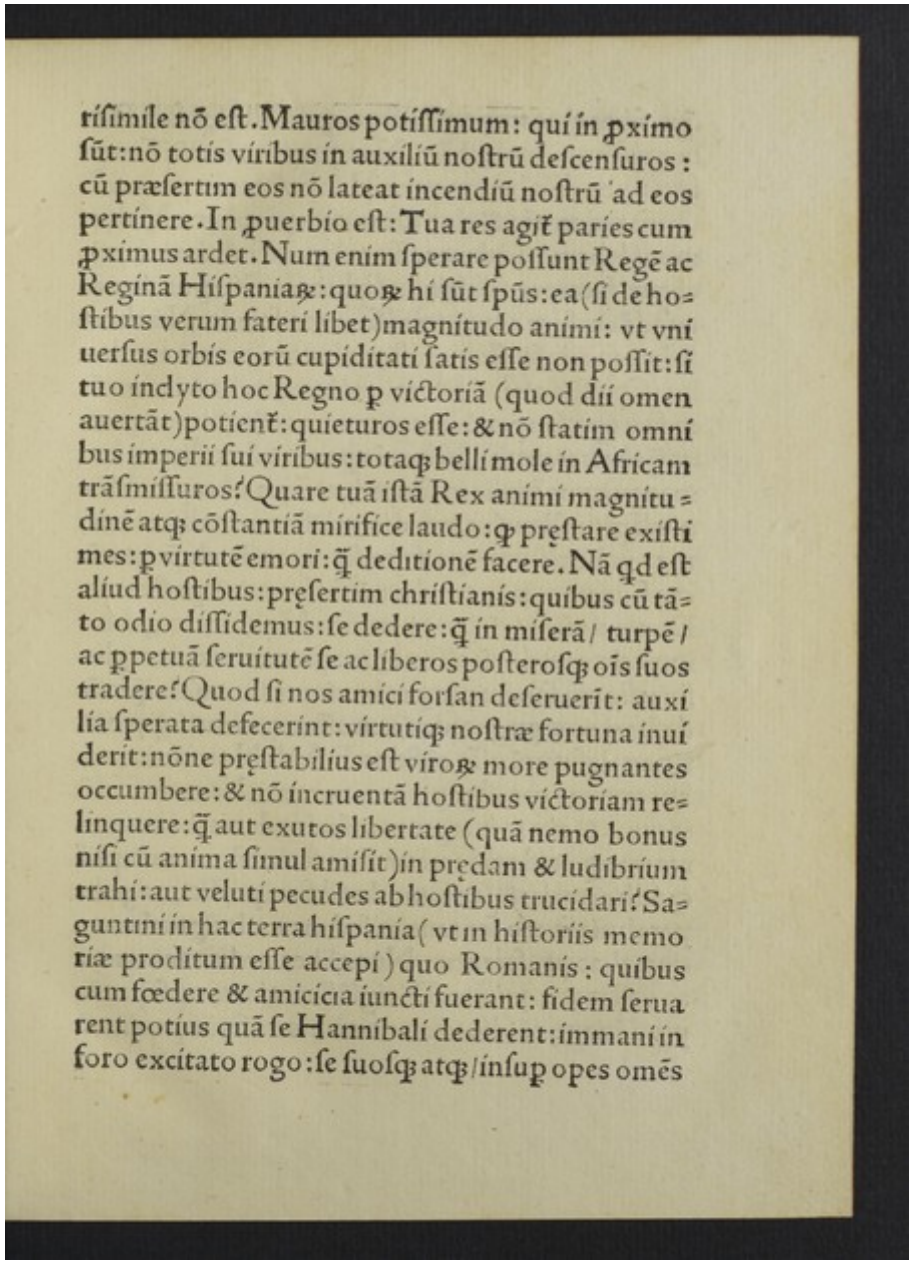
800 rechtshistorische Titel des 16. bis 18. Jhd. aus der Sammlung Desbillons werden digitalisiert und mit computerlinguistischen Verfahren analysiert.

Besonderheiten

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

Anforderungen

- Hohe Erkennungsgenauigkeit, insbesondere Wortgenauigkeit


rísímile nõ eſt. Mauros potíſſimum: quí ín ꝓxímo
ſũt: nõ totís víribus ín auxílíũ noſtrũ deſcenſuros:
cũ præſertim eos nõ lateat íncendíũ noſtrũ ad eos
pertínere. In ꝓuerbío eſt: Tua res agit paries cum
ꝓxímus ardet. Num enim ſperare poſſunt Regẽ ac
Reginã Híſpaniæ: quos hí ſũt ſpũs: ea (ſi de ho-
ſtíbus verum fateri líbet) magnítudo animí: vt vní
uerſus orbís eorũ cupíditatí ſatís eſſe non poſſit: ſí
tuo índyto hoc Regno ꝓ victoriã (quod díí omen
auertãt) potient: quíeturos eſſe: & nõ ſtatím omní
bus ímperíí ſuí víribus: totaq; bellí mole ín Africam
trãſmíſſuros? Quare tuã iſtã Rex animí magnítu-
dínẽ atq; cõſtantíã mírífice laudo: q; ꝓſtare exíſti
mes: ꝓ virtutẽ emorí: q̃ deditionẽ facere. Nã qd eſt
alíud hoſtíbus: ꝓſertim chriſtíanís: quíbus cũ tã-
to odio díſſídemus: ſe dedere: q̃ ín míſerã / turpẽ /
ac ꝓpetuã ſeruítutẽ ſe ac líberos poſteroſq; oĩs ſuos
tradere? Quod ſí nos amící forſan deſeruerít: auxí
lía ſperata defecerint: vírtutíq; noſtræ fortuna ínuí
derít: nõne ꝓſtabílíus eſt víros more pugnantes
occumbere: & nõ íncruentã hoſtíbus víctoríam re-
línquere: q̃ aut exutos líbertate (quã nemo bonus
níſí cũ anima ſímul amíſít) ín prędam & ludíbrium
trahí: aut velutí pecudes ab hoſtíbus trucídarí? Sa-
guntíní ín hac terra híſpanía (vt in híſtoríís memo
ríæ prodítum eſſe accepí) quo Romanís: quíbus
cum fœdere & amícícia íuncti fuerant: fídem ſerua
rent potíus quã ſe Hannibalí dederent: ímmaní ín
foro excítato rogo: ſe ſuoſq; atq; ínſuꝑ opes omẽs

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

B

Dividenden auf Stammaktien:
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)

Aus den Bilanzen (in 1 000 DM)

	1976	1977
Anlagevermögen	4 395	5 305
Umlaufvermögen	10 260	10 788
(flüssige Mittel)	519	727
Eigenkapital	5 117	5 117
(Grundkapital)	3 015	3 015
Fremdkapital	9 360	10 833
Bilanzgewinn	242	185
Bilanzsumme	14 655	16 094

Aus den Gewinn- und Verlustrechnungen

	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

BASF

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 Jentzsch, 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 Roßmöller, 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 Veredlungsstufen 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).

Deutscher Reichsanzeiger
und
Königlich Preußischer Staatsanzeiger.

№ 112. Berlin, Freitag, den 12. Mai, Abends. 1916.

Besonderheiten

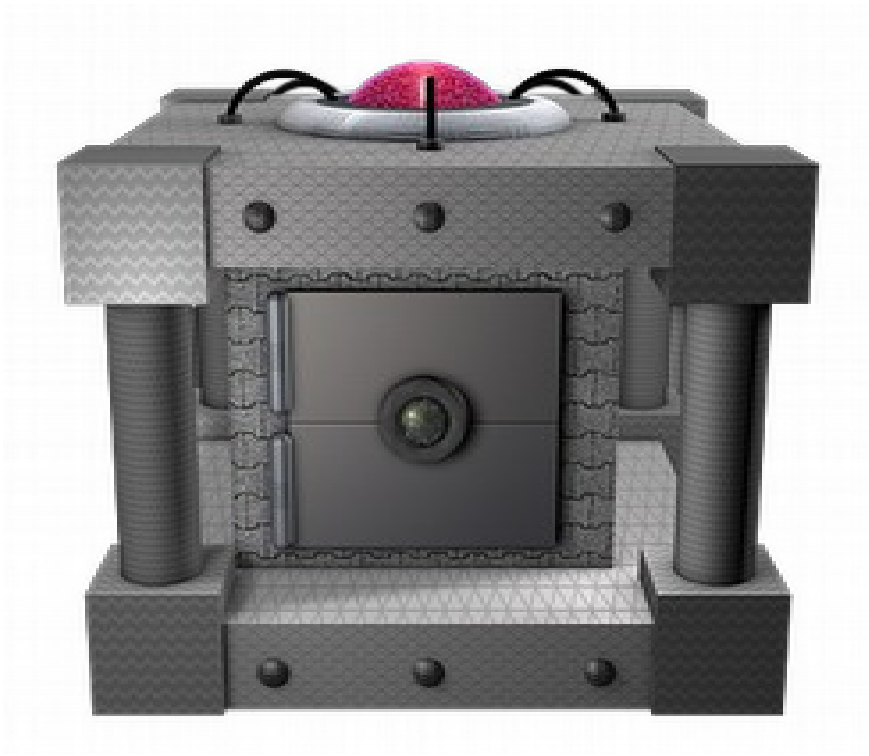
- Fraktur-Schrift
- Scans von Mikrofilmen in teilweise mäßiger Qualität
- Menge (127 Jahre, über 38000 Ausgaben, 25 TB TIFF Scans)

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/

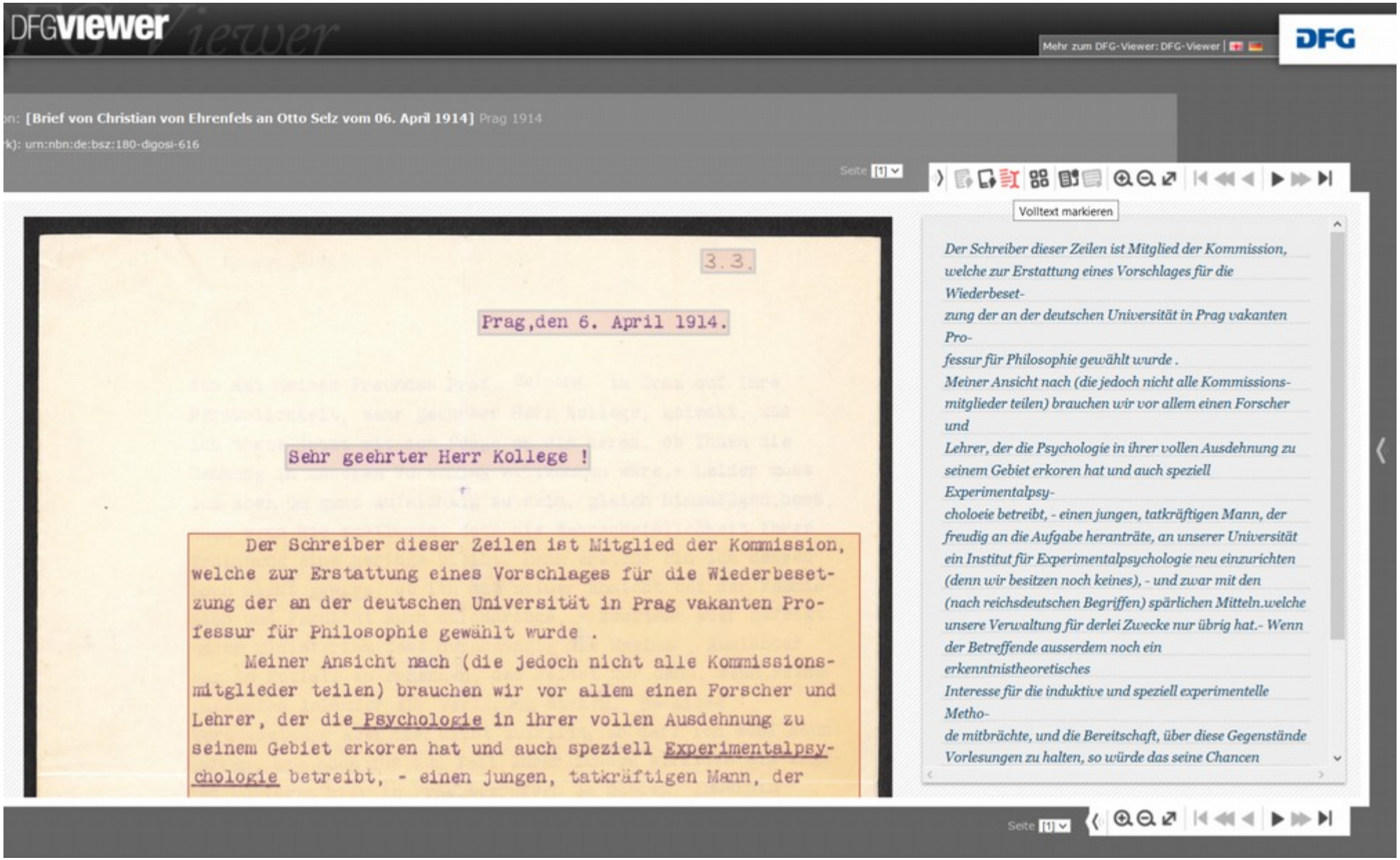
Tesseract

- 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



Workflow-Einbindung

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

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

```
$ abbyyocr11 -rl German -if max/275308_0089.jpg \
   -f PDF -of pdf/275308_0089.pdf \
   -f ALTO -of alto/275308_0089.xml \
   -f XML -of abbyy/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

Erkennungsgenauigkeit messen

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

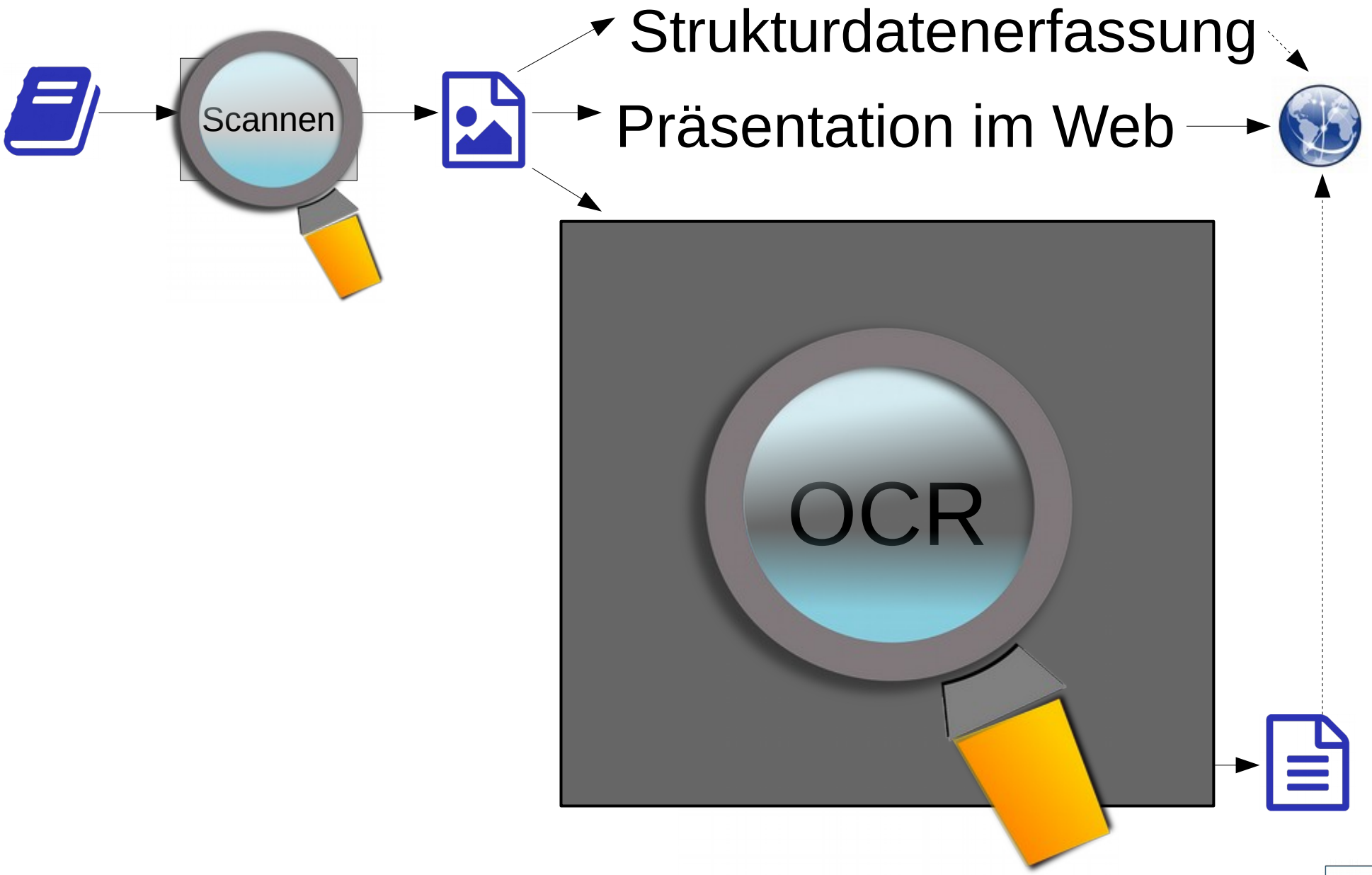
```
$ ocropus-errs *.gt.txt

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

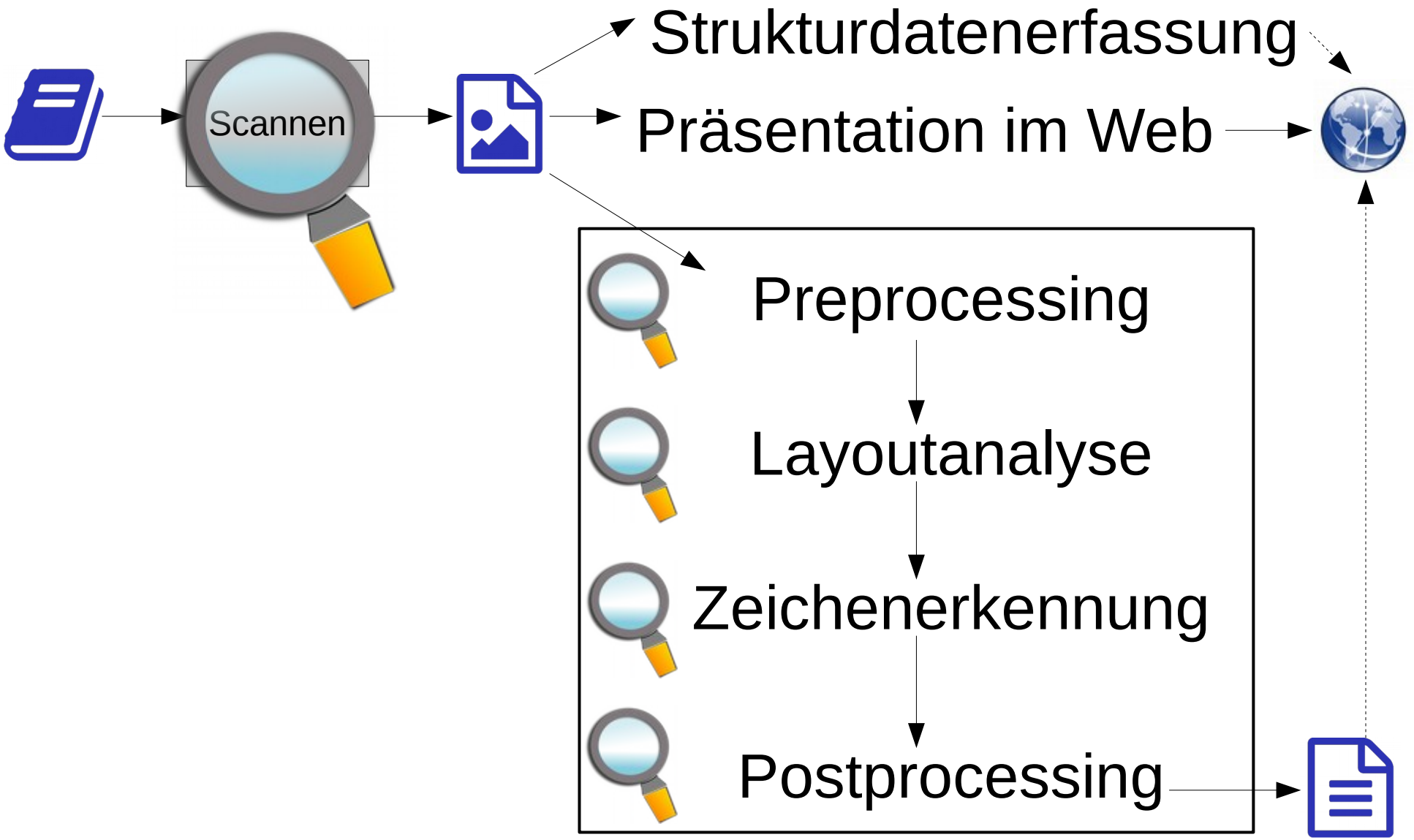
```
$ ocropus-econf *.gt.txt

1   _    -
1   W   w
1   H_  fi
1   !    _
1        _____  that
1   _____
1   H_  fl
1   e
1        _
1   _
```

Erkennungsgenauigkeit verbessern



Erkennungsgenauigkeit verbessern

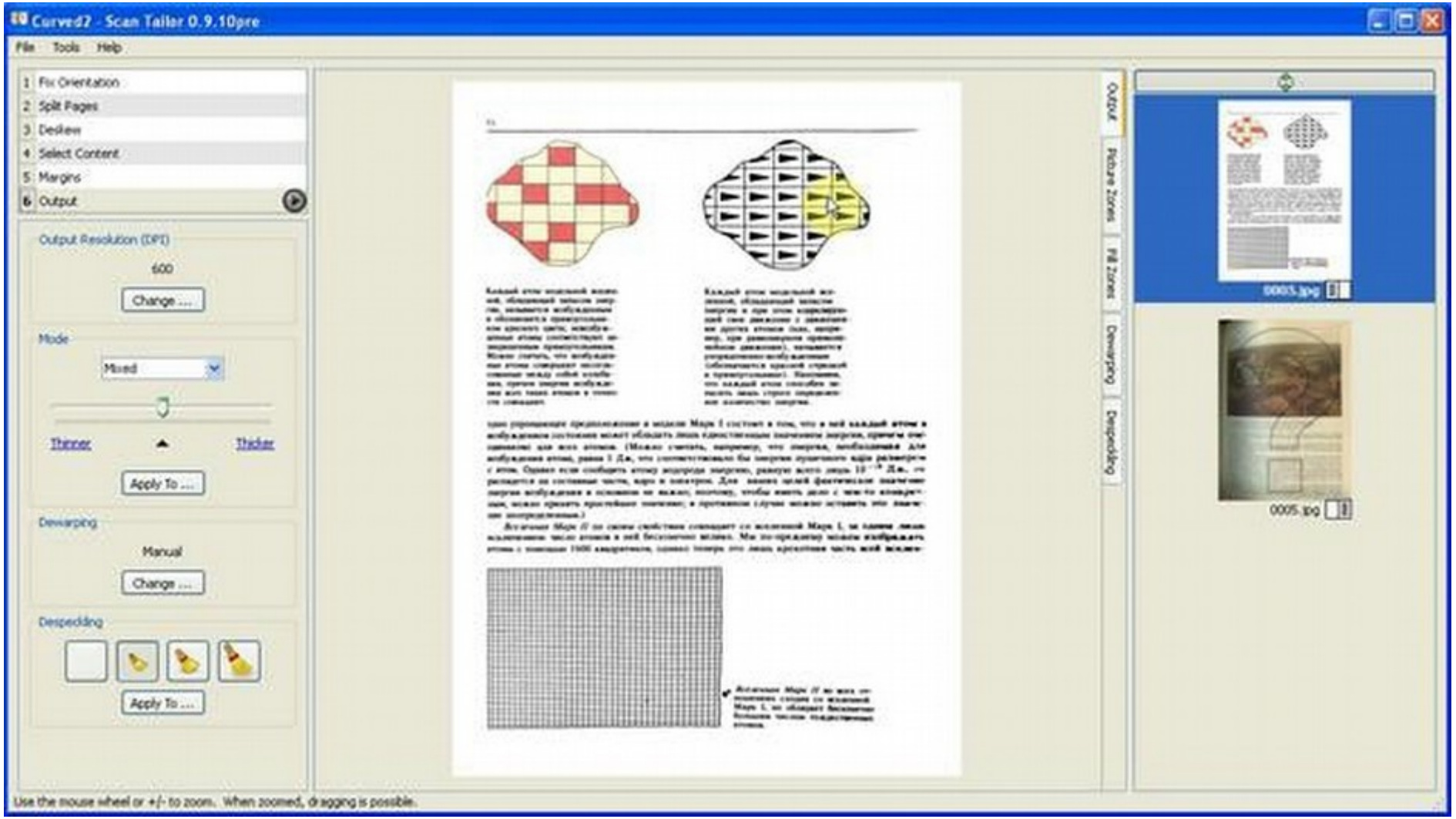


Preprocessing

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

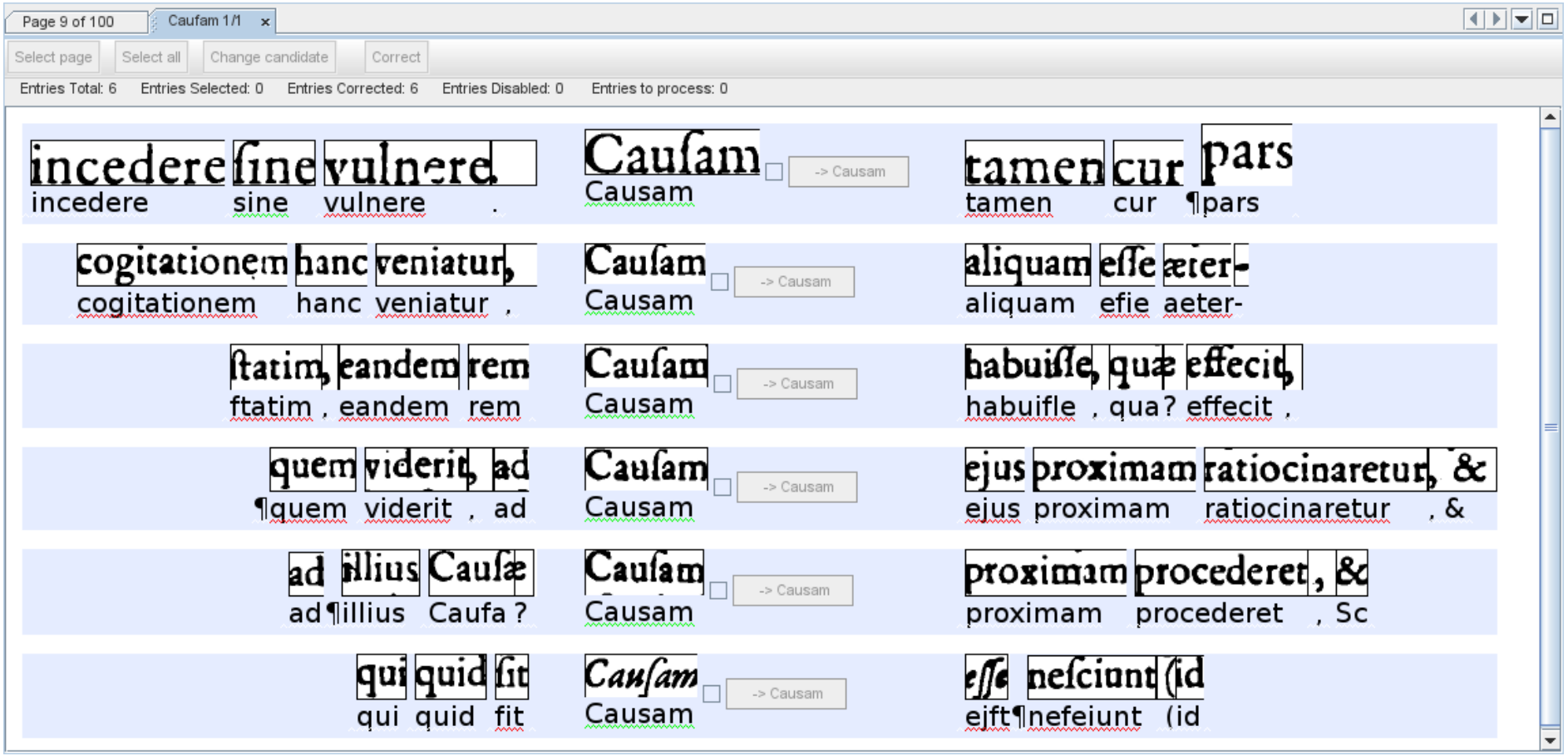
Preprocessing: Beispiel ScanTailor

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)



Resümee

- OCR: Keine One-Size-Fits-All-Lösungen
- Freie OCR Software sind konkurenzfähig 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)