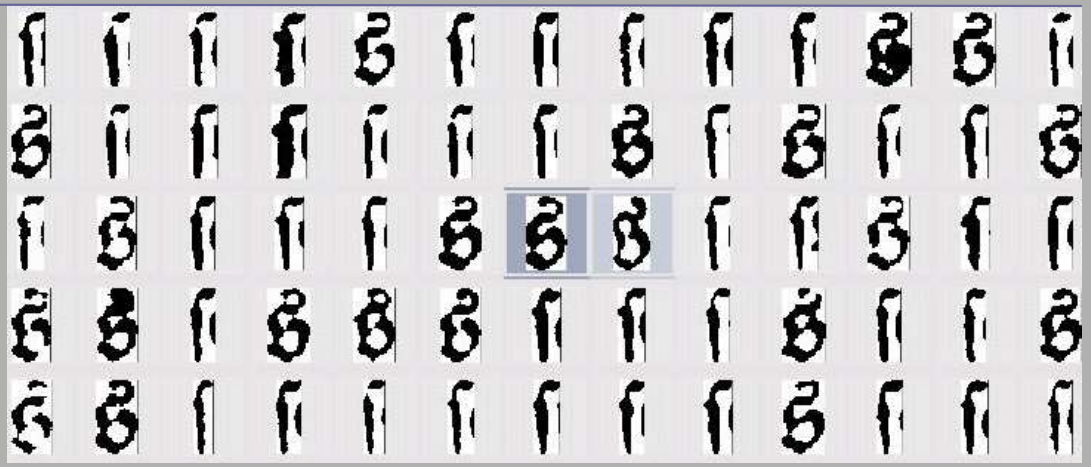
Improving Access to Text

IMPACT



IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands.

Verbesserter Zugang zu historischen Texten – Demonstration einiger Werkzeuge des Projekts IMPACT

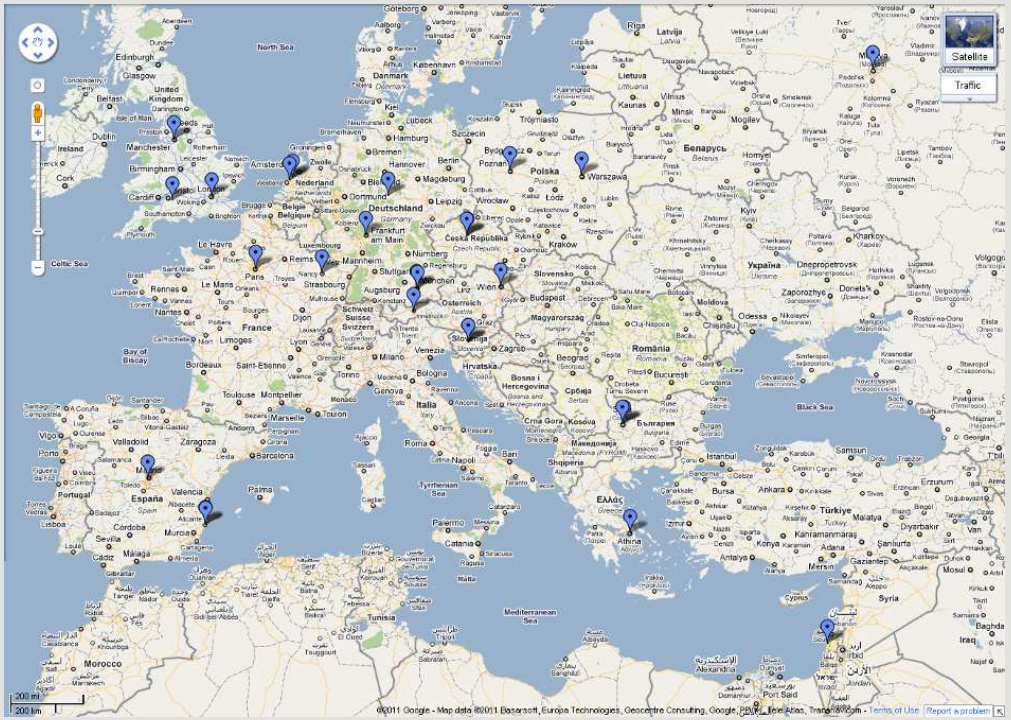
Mark-Oliver Fischer, Münchener Digitalisierungszentrum, Bayerische Staatsbibliothek

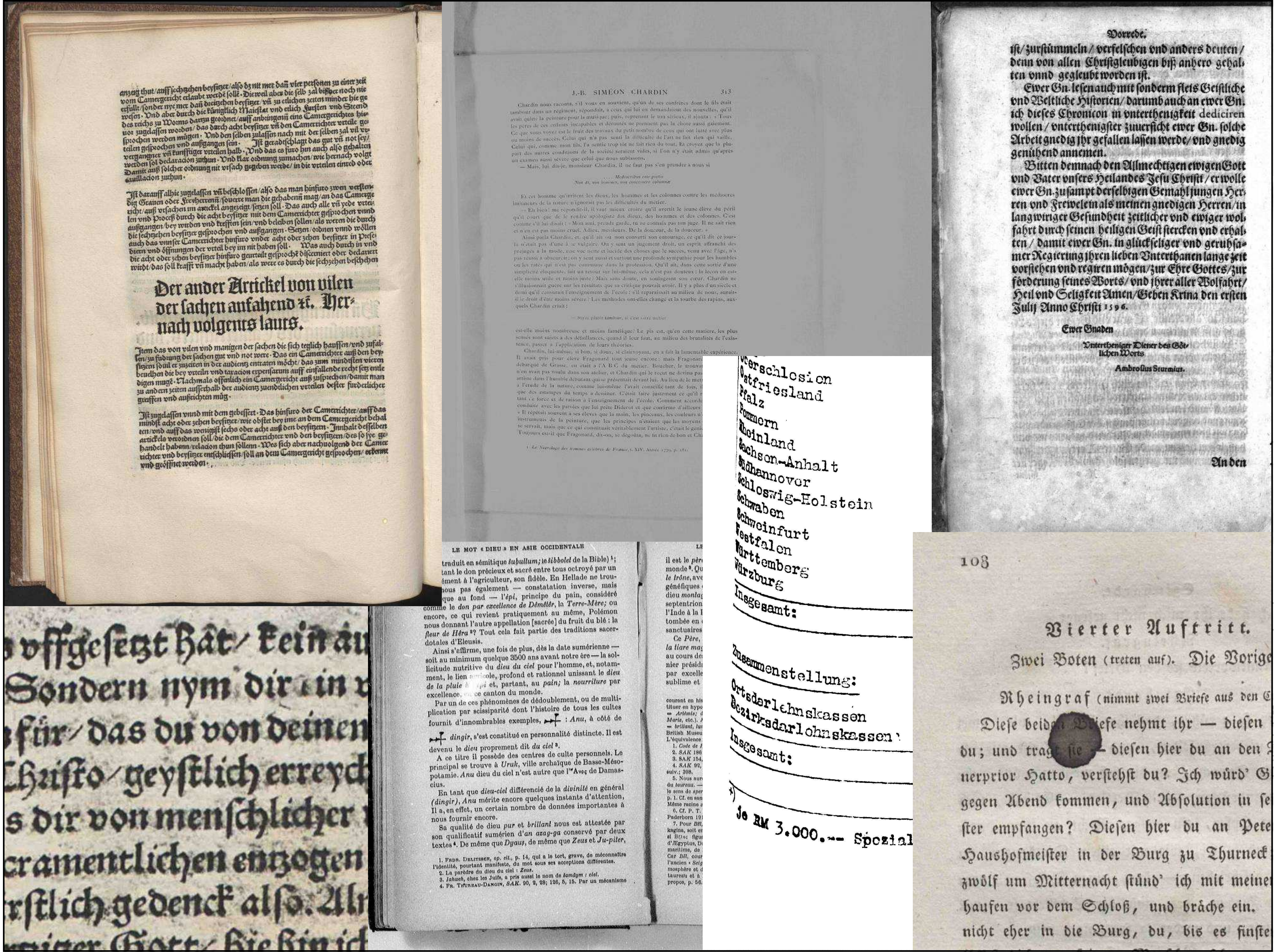
08.06.2011 - Deutscher Bibliothekartag,
Berlin

IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands.

IMPACT: “IMProving ACcess to Text”

- Ziel: Elektronischen Zugang zu historischen Texten verbessern
- Unter anderem leichtere, bessere Volltexterstellung: (Weiter-)Entwicklung von OCR-Software und Software rund um die eigentliche Texterkennung
- EU-gefördertes Kooperationsprojekt
- 26 Institutionen aus 13 Ländern: Bibliotheken (z.B. BSB, DNB, BL), Forschungsinstitutionen (z.B. NCSR Demokritos, INL), kommerzielle Partner (Abbyy, IBM)

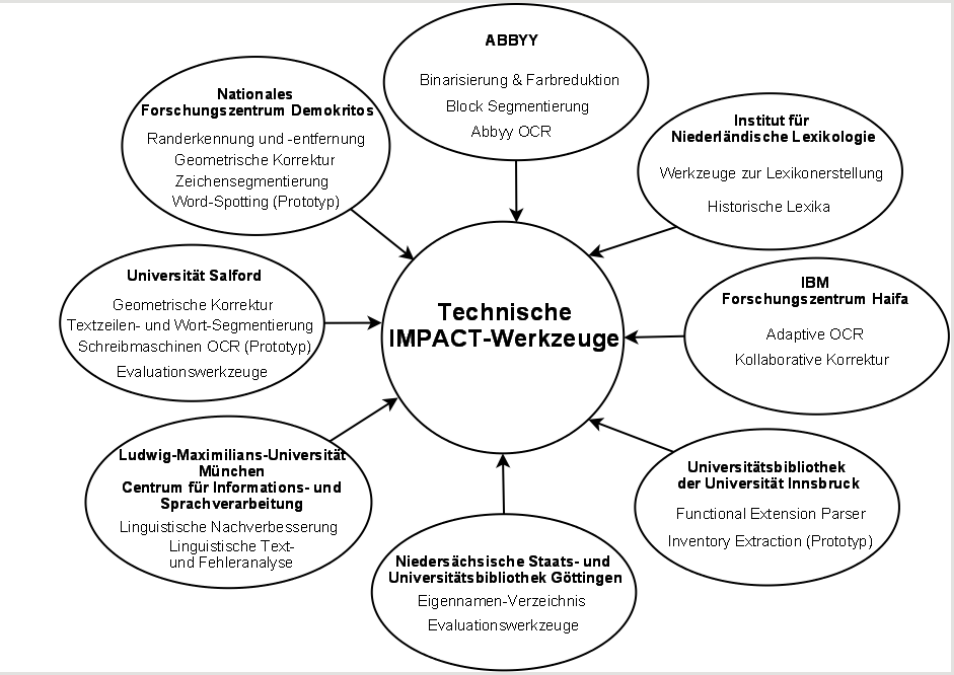



Der ander Artickel von vilen der sachen anfahend ꝛc. Hernach volgents lauts.
J.-B. SIMÉON CHARDIN 313
Vorrede.
ist/zurstümmeln/verfelschen und anders deuten/ denn von allen Christgleubigen biß anhero gehalten unnd gegleubt worden ist.
Ewer Gn. lesen auch mit sonderm fleis Geistliche und Weltliche Historien/ darumb auch an ewer Gn. ich dieses Chronicon in unterthenigkeit dedicirn wollen/ unterthenigster zuuersicht ewer Gn. solche Arbeit gnedig jhr gefallen lassen werde/ und gnedig genühend annemen.
Bitten demnach den Allmechtigen ewigen Gott und Vater unsers Heilandes Jesu Christi/ er wolle ewer Gn. zusampt derselbigen Gemahl jungen Herren und Frewelein als meinen gnedigen Herren/ in langwiriger Gesundheit zeitlicher und ewiger wolfahrt durch seinen heiligen Geist stercken und erhalten/ damit ewer Gn. in glückseliger und geruhsamer Regierung jhren lieben Unterthanen lange zeit vorstehen und regiren mögen/ zur Ehre Gottes/ zur förderung seines Worts/ und jhrer aller Wolfahrt/ Heil und Seligkeit Amen/ Geben Krina den ersten Julij Anno Christi 1596.
Ewer Gnaden
Unterthenigster Diener des Göttlichen Worts
Ambrosius Sturmius.
An den
LE MOT « DIEU » EN ASIE OCCIDENTALE
Oberschlesien
Ostfriesland
Pfalz
Pommern
Rheinland
Sachsen-Anhalt
Südhannover
Schleswig-Holstein
Schwaben
Schweinfurt
Westfalen
Württemberg
Würzburg
Insgesamt:
Zusammenstellung:
Ortsdarlehnskassen
Bezirksdarlehnskassen
Insgesamt:
*)
Je RM 3.000.-- Spezial
108
Vierter Auftritt.
Zwei Boten (treten auf). Die Vorige
Rheingraf (nimmt zwei Briefe aus den
Diese beiden Briefe nehmt ihr — diesen
du; und trag sie — diesen hier du an den
nerprior Hatto, verstehst du? Ich würd'
gegen Abend kommen, und Absolution in
ster empfangen? Diesen hier du an Pete
Haushofmeister in der Burg zu Thurneck
zwölf um Mitternacht stünd' ich mit meine
haufen vor dem Schloß, und bräche ein.
nicht eher in die Burg, du, bis es finste
vffgesetzt hat/ kein au
Sondern nym dir in
für/ das du von deinen
Christo/ geystlich erreych
s dir von menschlicher
cramentlichen entzogen
rstlich gedenck also. Al

IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands.

Tools für alle Prozessschritte eines OCR-Workflows

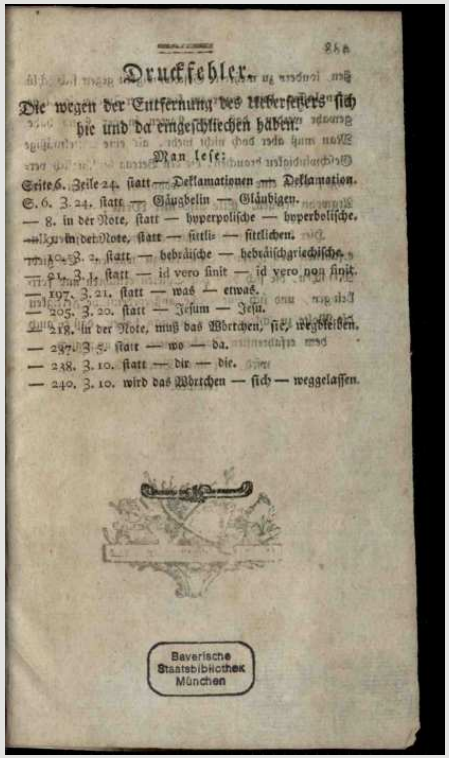
- Bildoptimierung
- Texterkennung
- Nachbearbeitung
- Nicht **eine** IMPACT-Software, sondern Zusammenspiel individueller Tools

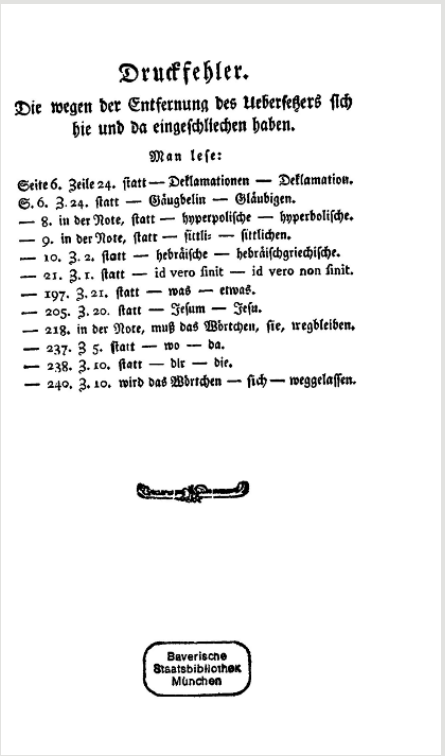


IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands.

Bildoptimierung

- Aufbereitung der Bilder für möglichst gute Texterkennung
- Nicht zwingend identisch mit optisch gutem Bild
- Tools zur geometrischen Korrektur: (NCSR Demokritos; Uni Salford)
  - Geraderücken von Seiten mit schiefem Textspiegel
  - Entzerren von Wellen, Bögen, … im Text
  - Trennung von Doppelseiten
  - Entfernung von Bildrändern
- Verbesserte, ‚intelligente' Binarisierung (Abbyy)





IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands.

Texterkennung

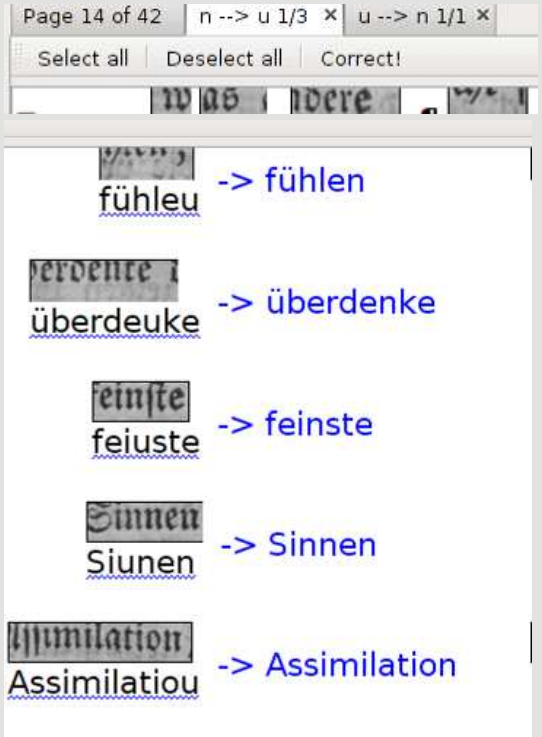
- In drei Schritten: Segmentierung, Zeichenerkennung, Wörterbuch
- IMPACT entwickelt Tools auf allen drei Ebenen
  - Verbesserte Segmentierung: Absätze, Wörter, Zeichen (Abbyy; NCSR Demokritos; Uni Salford)

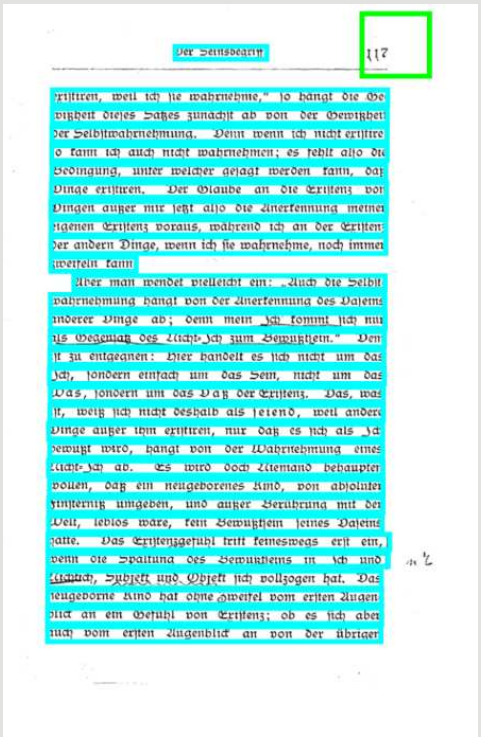
  - Deutliche Verbesserung der Fraktur-Erkennung (Abbyy)
  - Adaptive, selbstlernende OCR (IBM)
  - spezielle Maschinenschrift-OCR (Uni Salford, experimentell)
  - Spezifische Wörterbücher für alle Projektsprachen: Deutsch mit Schwerpunkt 16. Jahrhundert (Kooperation BSB und LMU München)

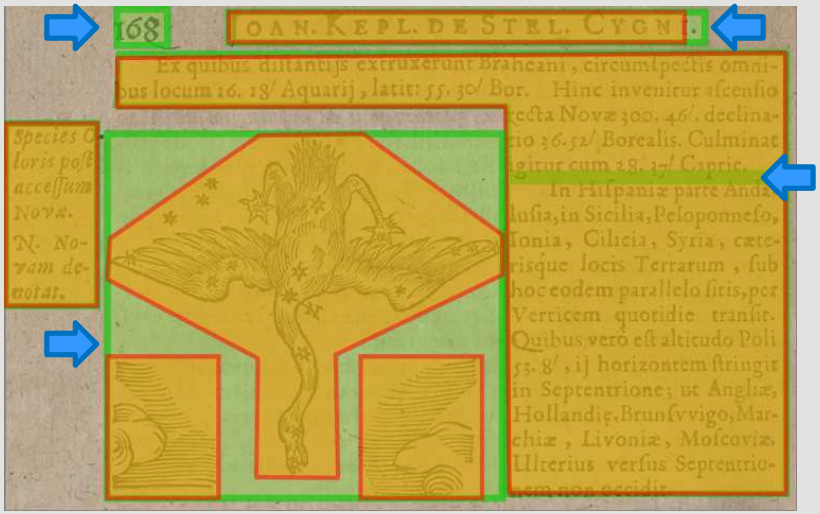
IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands.

Nachbearbeitung

- Software zur Nachkontrolle, Korrektur der OCR-Ergebnisse (IBM; LMU München)
- Strukturanalyse, -auszeichnung: Druckbereich, Fußnoten, Seitenzahlen, etc. (ULB Tirol)
- Tools zur automatischen Evaluation von OCR-Ergebnissen (Uni Salford; NCSR Demokritos)



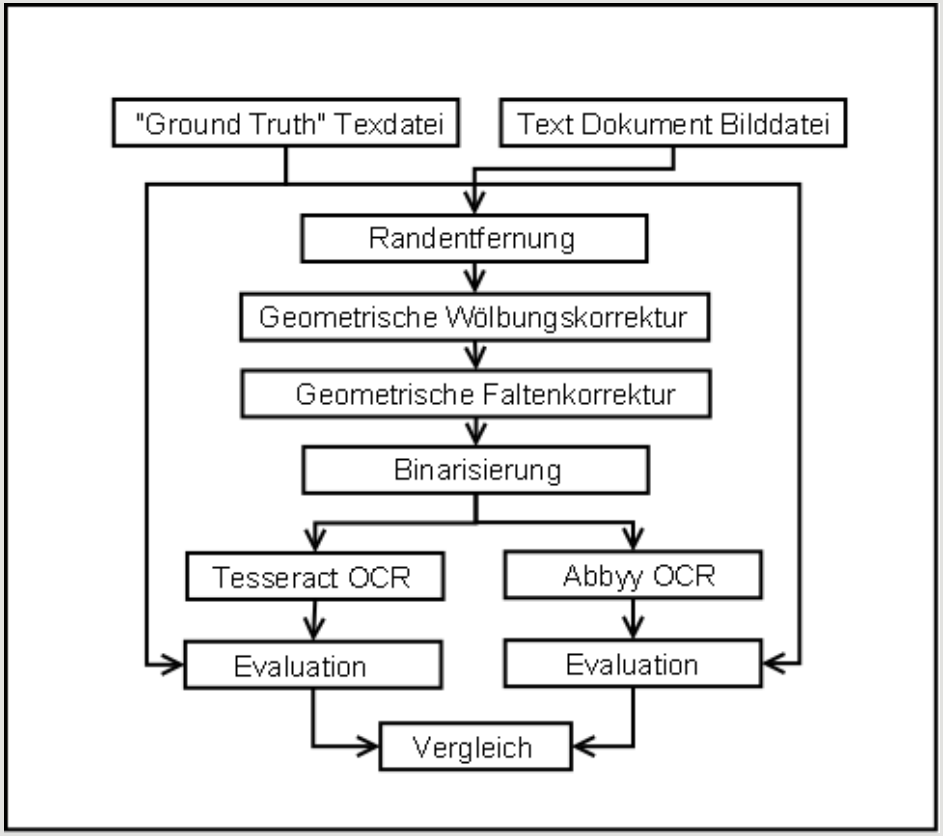




IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands.

Interoperabilität und Modularisierung

- Ziel: Alle IMPACT-Tools können zusammenwirken, sind miteinander einsetzbar
- Dazu stehen fast alle Tools als Webservices zur Verfügung, die zu individuellen ‚Workflows' kombiniert werden können



IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands.

Demonstration

- Stand der Deutschen Nationalbibliothek: G11
- Mittwoch und Donnerstag, je 13 bis 15 Uhr

- Weitere Informationen: http://www.impact-project.eu/