Kay-Michael Würzner

(Open-Source-)OCR-Workflows

Digital Humanities Kolloquium an der Berlin-Brandenburgischen Akademie der Wissenschaften ; (Berlin) : 2017.08.04

Persistent Identifier: urn:nbn:de:kobv:b4-opus-27863

Die vorliegende Datei wird Ihnen von der Berlin-Brandenburgischen Akademie der Wissenschaften unter einer Creative Commons Attribution- ShareAlike 3.0 Germany (cc by-sa 3.0) Licence zur Verfügung gestellt.

(Open-Source-)OCR-Workflows

Kay-Michael Würzner
wuerzner@bbaw.de

DH-Kolloquium an der Berlin-Brandenburgischen Akademie der Wissenschaften
4. August 2017

Übersicht

- Einleitung
  - Was ist OCR?
  - Wozu benutzt man OCR?
  - Warum überhaupt OCR?
- Technische Aspekte
  - Komponenten eines einfachen OCR-Workflows
  - Modelltraining
  - Optimierungsoptionen
  - Komplexere OCR-Workflows
- Nichttechnische Aspekte
  - OCR-D
  - Open-Source, und dann?

Was ist OCR?

Was ist OCR?

- **O**ptical **C**haracter **R**ecognition: Automatische Erfassung von Text in Bildern
- ursprünglich begrenzt auf Zeichenerkennung
- heute häufig Synonym für den gesamten Texterfassungsprozess
  - Bildvorverarbeitung
  - Layoutanalyse (OLR)
  - Zeilenerkennung
  - ...

"All the News That's Fit to Print."

The New York Times.

THE WEATHER

NEW YORK, MONDAY, NOVEMBER 11, 1918. TWENTY-FOUR PAGES.

ARMISTICE SIGNED, END OF THE WAR!
BERLIN SEIZED BY REVOLUTIONISTS;
NEW CHANCELLOR BEGS FOR ORDER;
OUSTED KAISER FLEES TO HOLLAND

SON FLEES WITH EX-KAISER

Hindenburg Also Believed to be Among Those in His Party.

ALL ARE HEAVILY ARMED

Automobiles Bristle with Rifles as Fugitives Arrive at Dutch Frontier.

ON THEIR WAY TO BE SHOT

Belgians Yell to Them, "Are You On Your Way to Paris?"

Kaiser Fought Hindenburg's Call for Abdication; Failed to Get Army's Support in Keeping Throne

Kaiser Skinned as He Signed Abdication

GERMAN DYN-STIES BEING WIPED OUT

King of Wuerttemberg Abdicates — Sovereign of Saxony to Follow Suit.

PRINCES MAY BE EXILED

Socialists Are Demanding That Every Sovereign in the Empire Shall be Dethroned.

BERLIN TROOPS JOIN REVOLT

Reds Shell Building in Which Officers Vainly Resist.

THRONES DEMAND REPUBLIC

Revolutionary Flag on Royal Palace—Crown Prince's Palace Also Seized.

GENERAL STRIKE IS BEGUN

Burgomaster and Police Submit—War Office Now Under Socialist Control.

MORE WARSHIPS JOIN THE REDS

Four Dreadnoughts in Kiel Harbor Espouse the Revolutionary Cause.

GUARDSHIPS ALSO GO OVER

Those Protecting Mines in the Great Belt and the Baltic Abandon Their Posts.

Socialist Chancellor Appeals to All Germans To Help Him Save Fatherland From Anarchy

WAR ENDS AT 6 O'CLOCK THIS MORNING

The State Department in Washington Made the Announcement at 2:45 o'Clock.

ARMISTICE WAS SIGNED IN FRANCE AT MIDNIGHT

Terms Include Withdrawal from Alsace-Lorraine, Disarming and Demobilization of Army and Navy, and Occupation of Strategic Naval and Military Points.

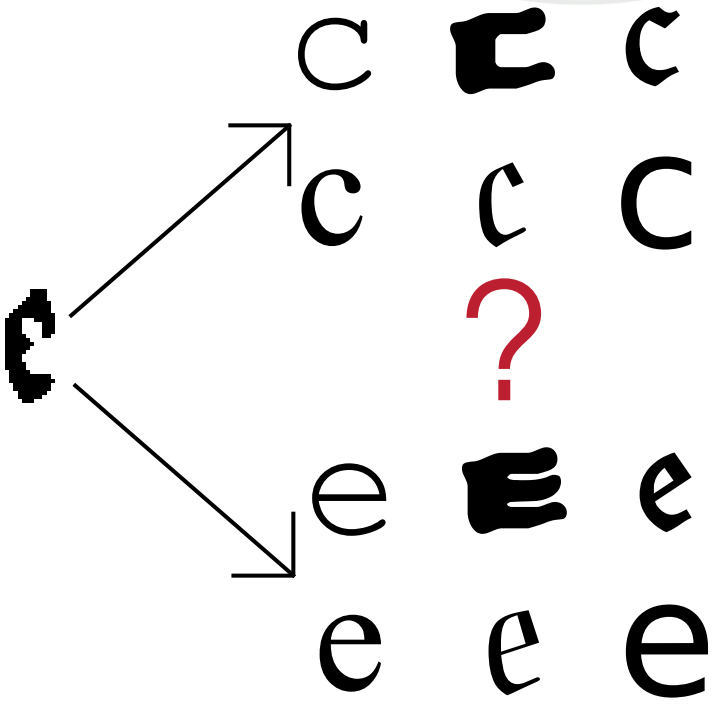
Zeichenorientierte Ansätze

- Erkennung erfolgt *glyphenweise*

  Pattern matching: Vergleich der Zeichenbilder zu in einem „Setzkasten“ gespeicherten Glyphen **Pixel für Pixel**

  Feature extraction: Zerlegung der Glyphen in vordefinierte, bedeutungstragende **Eigenschaften** wie *Einfärbung*, *Kurven*, *Linien* etc. und Vergleich zu Trainingsmaterialien

- Kombination beider Ansätze möglich
- Zerlegung der Seite in *Zeilen* und *Zeichen* notwendig
- Open-Source Software `Tesseract 3` *(Smith 2007)*
  - Einsatz verschiedener Lexika für frequente Wörter, Sonderzeichen und häufige Fehler zur Verbesserung der Erkennung

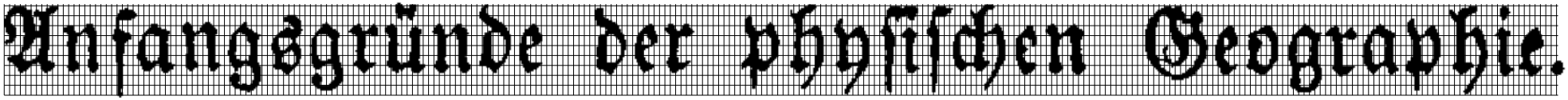


Zeilenorientierte Ansätze

- Erkennung erfolgt *zeilenweise*

  Scaling: Einheitliche Höhe für alle Zeilen

  Feature extraction: Grid mit festgelegter Anzahl (horizontaler) Zeilen und variabler Anzahl (vertikaler) Spalten: Zeilen als Sequenzen binärwertiger Vektoren fixer Länge



- Kontextsensitive (i.e. über *Übergangswahrscheinlichkeiten* der Vektoren) Erkennung
- Zerlegung der Seite in *Zeilen* notwendig
- Vorgehen (normalerweise) *robuster* gegenüber Varianz durch Artefakte als zeichenorientierte Ansätze
- Open-Source Software `OCRopus` *(Breuel 2008)*
  - Einsatz *neuronaler Netzwerke* für die Sequenzklassifikation

Wozu braucht man OCR?

Wozu braucht man OCR?

- typische Anwendungen:
  - Nummernschilderkennung (*Automatic number plate recognition*)

Image by Achim Raschka, CC BY-SA 3.0

Wozu braucht man OCR?

- typische Anwendungen:
  - Nummernschilderkennung (*Automatic number plate recognition*)
  - Captcha-Umgehung (*Completely Automated Public Turing test to tell Computers and Humans Apart*)

Image by JD, CC BY-SA 2.0

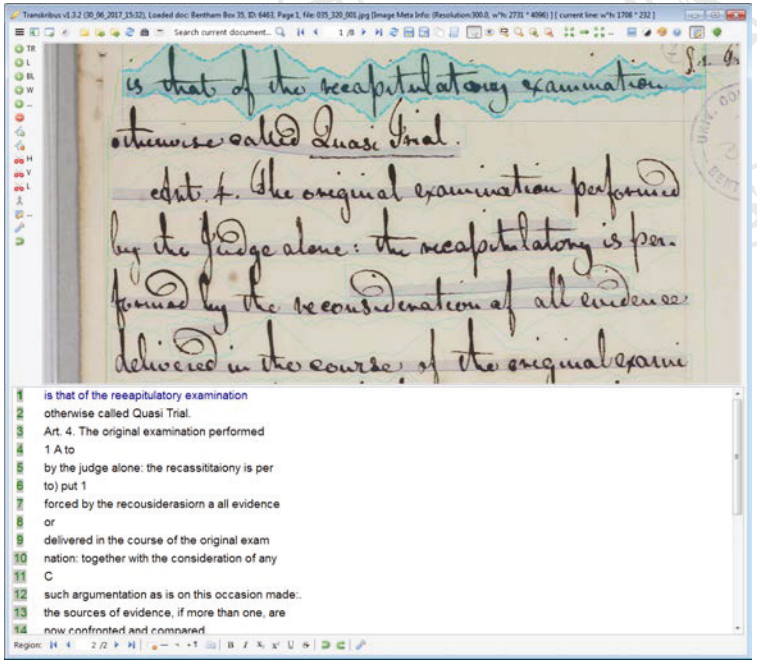
Wozu braucht man OCR?

- typische Anwendungen:
  - Nummernschilderkennung (*Automatic number plate recognition*)
  - Captcha-Umgehung (*Completely Automated Public Turing test to tell Computers and Humans Apart*)
  - Schlüsselinformationsextraktion (*Document's key information extraction*)

Image by Eluminary, CC BY-SA 4.0

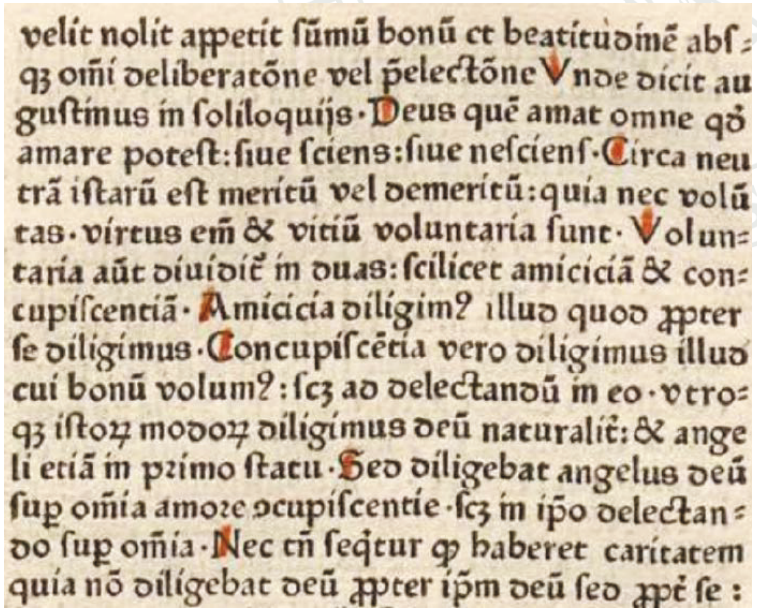
Wozu braucht man OCR?

- typische Anwendungen:
  - Nummernschilderkennung (*Automatic number plate recognition*)
  - Captcha-Umgehung (*Completely Automated Public Turing test to tell Computers and Humans Apart*)
  - Schlüsselinformationsextraktion (*Document's key information extraction*)
  - Handschrifterkennung (*Handwritten text recognition*)



Wozu braucht man OCR?

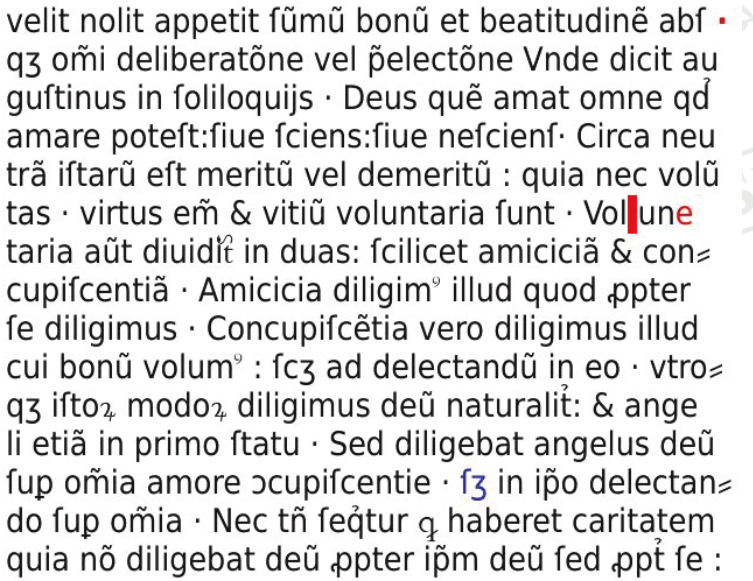
- typische Anwendungen:
  - Nummernschilderkennung (*Automatic number plate recognition*)
  - Captcha-Umgehung (*Completely Automated Public Turing test to tell Computers and Humans Apart*)
  - Schlüsselinformationsextraktion (*Document's key information extraction*)
  - Handschrifterkennung (*Handwritten text recognition*)
  - **Volltextdigitalisierung**


velit nolit appetit ſũmũ bonũ et beatitudinẽ abſ=
q3 oñi deliberatõne vel ꝑelectõne Vnde dicit au
guſtinus in ſoliloquijs·Deus quẽ amat omne qd
amare poteſt: ſiue ſciens: ſiue neſcienſ·Circa neu
trã iſtarũ eſt meritũ vel demeritũ: quia nec volũ
tas·virtus eñ & vitiũ voluntaria ſunt·Volun=
taria aũt diuidit in duas: ſcilicet amicicią & con=
cupiſcentią·Amicicia diligim? illud quod ꝓpter
ſe diligimus·Concupiſcẽtia vero diligimus illud
cui bonũ volum?: ſcz ad delectandũ in eo·vtro=
q3 iſtoꝝ modoꝝ diligimus deũ naturaliť: & ange
li etiã in primo ſtatu·Sed diligebat angelus deũ
ſup oñia amore ꝯcupiſcentie·ſcz in ipo delectan=
do ſup oñia·Nec tñ ſequtur q3 haberet caritatem
quia nõ diligebat deũ ꝓpter ipm deũ ſed ꝓꝑ ſe:

Images by Uwe Springmann, CC BY-SA 4.0

Wozu braucht man OCR?

- typische Anwendungen:
  - Nummernschilderkennung (*Automatic number plate recognition*)
  - Captcha-Umgehung (*Completely Automated Public Turing test to tell Computers and Humans Apart*)
  - Schlüsselinformationsextraktion (*Document's key information extraction*)
  - Handschrifterkennung (*Handwritten text recognition*)
  - **Volltextdigitalisierung**


velit nolit appetit ſũmũ bonũ et beatitudinẽ abſ ·
qȝ om̃i deliberatõne vel p̃electõne Vnde dicit au
guſtinus in ſoliloquijs · Deus quẽ amat omne qd̓
amare poteſt:ſiue ſciens:ſiue neſcienſ· Circa neu
trã iſtarũ eſt meritũ vel demeritũ : quia nec volũ
tas · virtus em̃ & vitiũ voluntaria ſunt · Vol une
taria aũt diuidit in duas: ſcilicet amiciciã & con≠
cupiſcentiã · Amicicia diligim⁹ illud quod ꝓpter
ſe diligimus · Concupiſcẽtia vero diligimus illud
cui bonũ volum⁹ : ſcȝ ad delectandũ in eo · vtro≠
qȝ iſto₄ modo₄ diligimus deũ naturaliť: & ange
li etiã in primo ſtatu · Sed diligebat angelus deũ
ſup om̃ia amore ↄcupiſcentie · ſȝ in ip̃o delectan≠
do ſup om̃ia · Nec tñ ſeq̓tur q̱ haberet caritatem
quia nõ diligebat deũ ꝓpter ip̃m deũ ſed ꝓpť ſe :

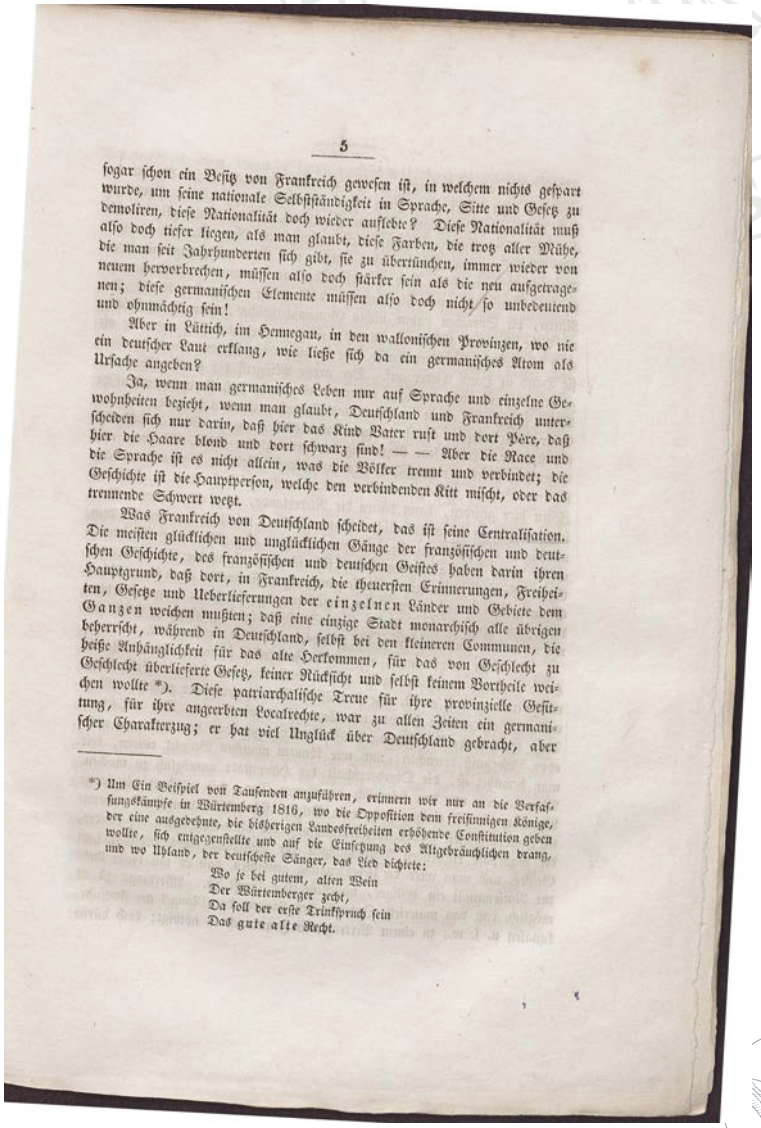
Images by Uwe Springmann, CC BY-SA 4.0

Warum überhaupt OCR?

- OCR ist immer **fehlerhaft**! Aber:
- verändertes „Rechercheverhalten“ in Zeiten zunehmender Verfügbarkeit digitaler Quellen
  - Wissenserwerb durch Internetsuche
  - Sekundärliteratur (fast) vollständig **textdigital** verfügbar
  - Navigationssystem vs. Autoatlas
- Ansprüche an Verfügbarkeit von **Primärquellen** wächst
- vielfältige quantitative Auswertungsmethoden (i.e. *distant reading*)
- für den **Digital Humanist**: Bruch mit dem „Diktat der Verfügbarkeit“

Komponenten eines einfachen OCR-Workflows

Komponenten eines einfachen OCR-Workflows



5

sogar schon ein Besitz von Frankreich gewesen ist, in welchem nichts gespart wurde, um seine nationale Selbstständigkeit in Sprache, Sitte und Gesetz zu demoliren, diese Nationalität doch wieder auflebte? Diese Nationalität muß also doch tiefer liegen, als man glaubt, diese Farben, die trotz aller Mühe, die man seit Jahrhunderten sich gibt, sie zu übertünchen, immer wieder von neuem hervorbrechen, müssen also doch stärker sein als die neu aufgetragenen; diese germanischen Elemente müssen also doch nicht so unbedeutend und ohnmächtig sein!

Aber in Lüttich, im Hennegau, in den wallonischen Provinzen, wo nie ein deutscher Laut erklang, wie ließe sich da ein germanisches Atom als Ursache angeben?

Ja, wenn man germanisches Leben nur auf Sprache und einzelne Gewohnheiten bezieht, wenn man glaubt, Deutschland und Frankreich unterscheiden sich nur darin, daß hier das Kind Vater ruft und dort Père, daß hier die Haare blond und dort schwarz sind! — — Aber die Race und die Sprache ist es nicht allein, was die Völker trennt und verbindet; die Geschichte ist die Hauptperson, welche den verbindenden Kitt mischt, oder das trennende Schwert wetzt.

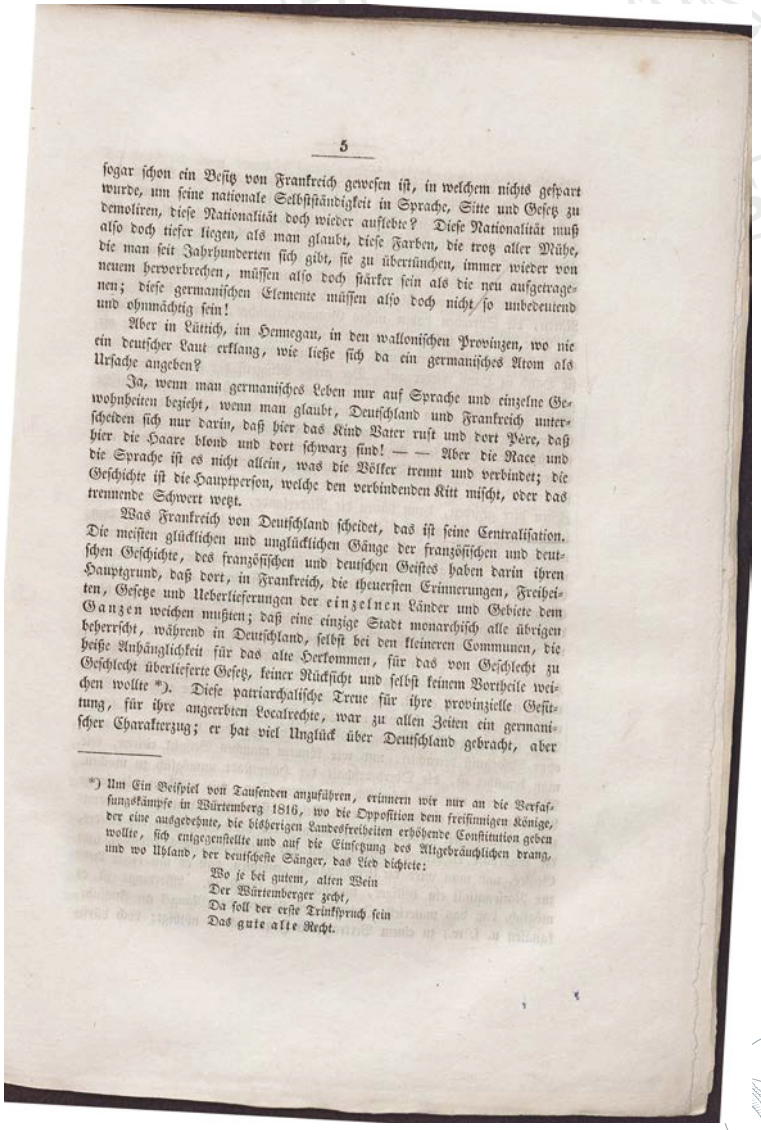
Was Frankreich von Deutschland scheidet, das ist seine Centralisation. Die meisten glücklichen und unglücklichen Gänge der französischen und deutschen Geschichte, des französischen und deutschen Geistes haben darin ihren Hauptgrund, daß dort, in Frankreich, die theuersten Erinnerungen, Freiheiten, Gesetze und Ueberlieferungen der einzelnen Länder und Gebiete dem Ganzen weichen mußten; daß eine einzige Stadt monarchisch alle übrigen beherrscht, während in Deutschland, selbst bei den kleineren Communen, die heiße Anhänglichkeit für das alte Herkommen, für das von Geschlecht zu Geschlecht überlieferte Gesetz, keiner Rücksicht und selbst keinem Vortheile weichen wollte *). Diese patriarchalische Treue für ihre provinzielle Gesittung, für ihre angeerbten Localrechte, war zu allen Zeiten ein germanischer Charakterzug; er hat viel Unglück über Deutschland gebracht, aber

*) Um Ein Beispiel von Tausenden anzuführen, erinnern wir nur an die Verfassungskämpfe in Würtemberg 1816, wo die Opposition dem freisinnigen Könige, der eine ausgedehnte, die bisherigen Landesfreiheiten erhöhende Constitution geben wollte, sich entgegenstellte und auf die Einsetzung des Altgebräuchlichen drang, und wo Uhland, der deutscheste Sänger, das Lied dichtete:

Wo je bei gutem, alten Wein
Der Würtemberger zecht,
Da soll der erste Trinkspruch sein
Das gute alte Recht.

Komponenten eines einfachen OCR-Workflows

1. Bildvorverarbeitung
2.
3.



5

sogar schon ein Besitz von Frankreich gewesen ist, in welchem nichts gespart wurde, um seine nationale Selbstständigkeit in Sprache, Sitte und Gesetz zu demoliren, diese Nationalität doch wieder auflebte? Diese Nationalität muß also doch tiefer liegen, als man glaubt, diese Farben, die trotz aller Mühe, die man seit Jahrhunderten sich gibt, sie zu übertünchen, immer wieder von neuem hervorbrechen, müssen also doch stärker sein als die neu aufgetragenen; diese germanischen Elemente müssen also doch nicht so unbedeutend und ohnmächtig sein!

Aber in Lüttich, im Hennegau, in den wallonischen Provinzen, wo nie ein deutscher Laut erklang, wie ließe sich da ein germanisches Atom als Ursache angeben?

Ja, wenn man germanisches Leben nur auf Sprache und einzelne Gewohnheiten bezieht, wenn man glaubt, Deutschland und Frankreich unterscheiden sich nur darin, daß hier das Kind Vater ruft und dort Père, daß hier die Haare blond und dort schwarz sind! — — Aber die Race und die Sprache ist es nicht allein, was die Völker trennt und verbindet; die Geschichte ist die Hauptperson, welche den verbindenden Kitt mischt, oder das trennende Schwert wetzt.

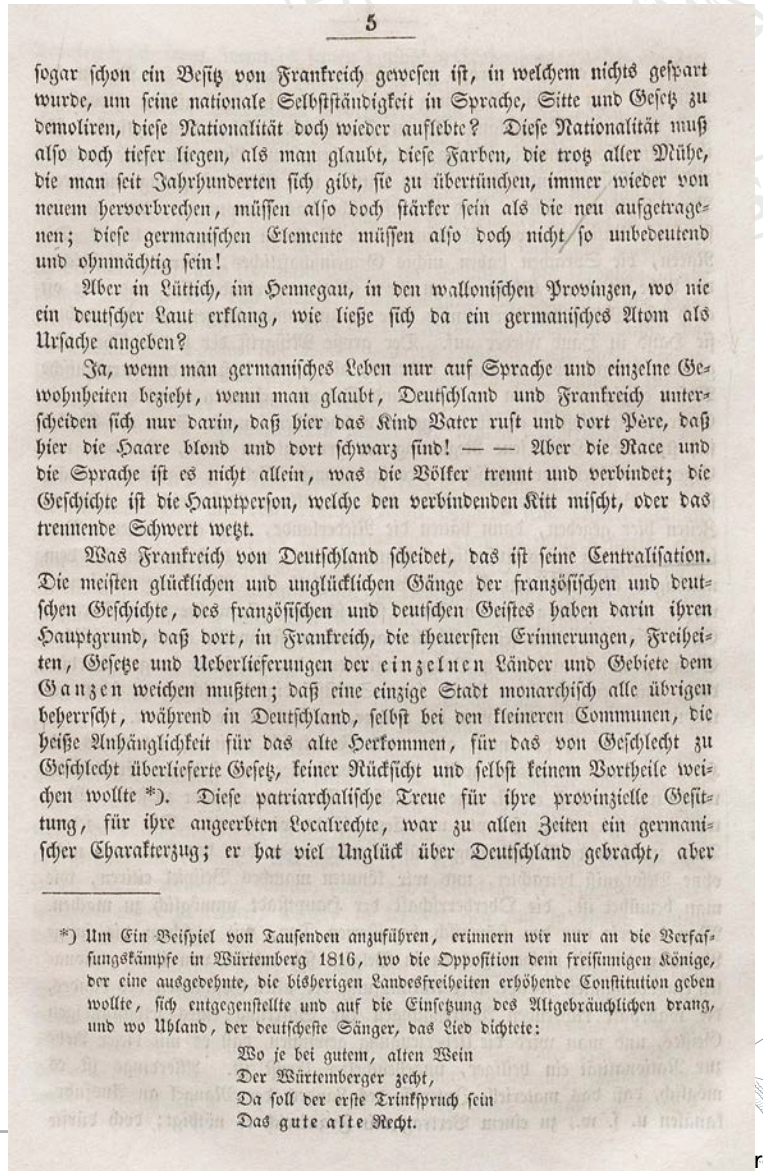
Was Frankreich von Deutschland scheidet, das ist seine Centralisation. Die meisten glücklichen und unglücklichen Gänge der französischen und deutschen Geschichte, des französischen und deutschen Geistes haben darin ihren Hauptgrund, daß dort, in Frankreich, die theuersten Erinnerungen, Freiheiten, Gesetze und Ueberlieferungen der einzelnen Länder und Gebiete dem Ganzen weichen mußten; daß eine einzige Stadt monarchisch alle übrigen beherrscht, während in Deutschland, selbst bei den kleineren Communen, die heiße Anhänglichkeit für das alte Herkommen, für das von Geschlecht zu Geschlecht überlieferte Gesetz, keiner Rücksicht und selbst keinem Vortheile weichen wollte *). Diese patriarchalische Treue für ihre provinzielle Gesittung, für ihre angeerbten Localrechte, war zu allen Zeiten ein germanischer Charakterzug; er hat viel Unglück über Deutschland gebracht, aber

*) Um Ein Beispiel von Tausenden anzuführen, erinnern wir nur an die Verfassungskämpfe in Würtemberg 1816, wo die Opposition dem freisinnigen Könige, der eine ausgedehnte, die bisherigen Landesfreiheiten erhöhende Constitution geben wollte, sich entgegenstellte und auf die Einsetzung des Altgebräuchlichen drang, und wo Uhland, der deutscheste Sänger, das Lied dichtete:

Wo je bei gutem, alten Wein
Der Würtemberger zecht,
Da soll der erste Trinkspruch sein
Das gute alte Recht.

Komponenten eines einfachen OCR-Workflows

1. Bildvorverarbeitung
2.
3.



5

sogar schon ein Besitz von Frankreich gewesen ist, in welchem nichts gespart wurde, um seine nationale Selbstständigkeit in Sprache, Sitte und Gesetz zu demoliren, diese Nationalität doch wieder auflebte? Diese Nationalität muß also doch tiefer liegen, als man glaubt, diese Farben, die trotz aller Mühe, die man seit Jahrhunderten sich gibt, sie zu übertünchen, immer wieder von neuem hervorbrechen, müssen also doch stärker sein als die neu aufgetragenen; diese germanischen Elemente müssen also doch nicht so unbedeutend und ohnmächtig sein!

Aber in Lüttich, im Hennegau, in den wallonischen Provinzen, wo nie ein deutscher Laut erklang, wie ließe sich da ein germanisches Atom als Ursache angeben?

Ja, wenn man germanisches Leben nur auf Sprache und einzelne Gewohnheiten bezieht, wenn man glaubt, Deutschland und Frankreich unterscheiden sich nur darin, daß hier das Kind Vater ruft und dort Père, daß hier die Haare blond und dort schwarz sind! — — Aber die Race und die Sprache ist es nicht allein, was die Völker trennt und verbindet; die Geschichte ist die Hauptperson, welche den verbindenden Kitt mischt, oder das trennende Schwert wetzt.

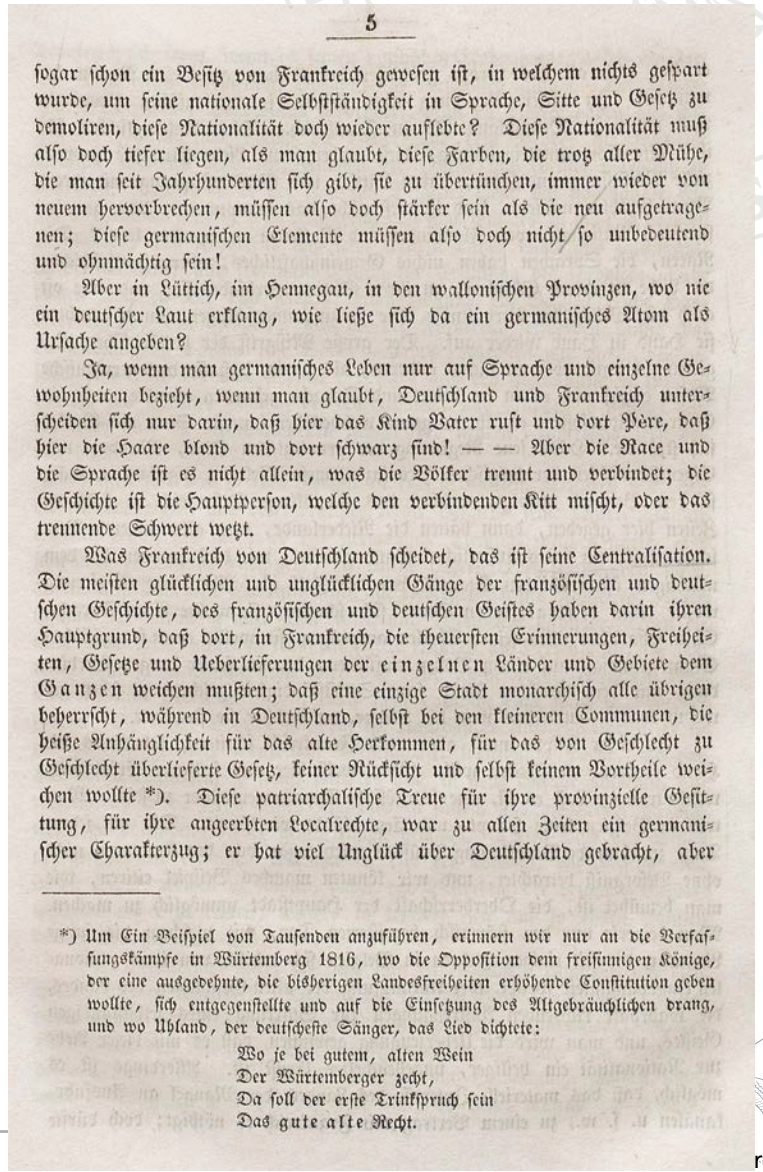
Was Frankreich von Deutschland scheidet, das ist seine Centralisation. Die meisten glücklichen und unglücklichen Gänge der französischen und deutschen Geschichte, des französischen und deutschen Geistes haben darin ihren Hauptgrund, daß dort, in Frankreich, die theuersten Erinnerungen, Freiheiten, Gesetze und Ueberlieferungen der einzelnen Länder und Gebiete dem Ganzen weichen mußten; daß eine einzige Stadt monarchisch alle übrigen beherrscht, während in Deutschland, selbst bei den kleineren Communen, die heiße Anhänglichkeit für das alte Herkommen, für das von Geschlecht zu Geschlecht überlieferte Gesetz, keiner Rücksicht und selbst keinem Vortheile weichen wollte *). Diese patriarchalische Treue für ihre provinzielle Gesittung, für ihre angeerbten Localrechte, war zu allen Zeiten ein germanischer Charakterzug; er hat viel Unglück über Deutschland gebracht, aber

*) Um Ein Beispiel von Tausenden anzuführen, erinnern wir nur an die Verfassungskämpfe in Würtemberg 1816, wo die Opposition dem freisinnigen Könige, der eine ausgedehnte, die bisherigen Landesfreiheiten erhöhende Constitution geben wollte, sich entgegenstellte und auf die Einsetzung des Altgebräuchlichen drang, und wo Uhland, der deutscheste Sänger, das Lied dichtete:

Wo je bei gutem, alten Wein
Der Würtemberger zecht,
Da soll der erste Trinkspruch sein
Das gute alte Recht.

Komponenten eines einfachen OCR-Workflows

1. Bildvorverarbeitung
2. Layoutanalyse
3.



5

sogar schon ein Besitz von Frankreich gewesen ist, in welchem nichts gespart wurde, um seine nationale Selbstständigkeit in Sprache, Sitte und Gesetz zu demoliren, diese Nationalität doch wieder auflebte? Diese Nationalität muß also doch tiefer liegen, als man glaubt, diese Farben, die trotz aller Mühe, die man seit Jahrhunderten sich gibt, sie zu übertünchen, immer wieder von neuem hervorbrechen, müssen also doch stärker sein als die neu aufgetragenen; diese germanischen Elemente müssen also doch nicht so unbedeutend und ohnmächtig sein!

Aber in Lüttich, im Hennegau, in den wallonischen Provinzen, wo nie ein deutscher Laut erklang, wie ließe sich da ein germanisches Atom als Ursache angeben?

Ja, wenn man germanisches Leben nur auf Sprache und einzelne Gewohnheiten bezieht, wenn man glaubt, Deutschland und Frankreich unterscheiden sich nur darin, daß hier das Kind Vater ruft und dort Père, daß hier die Haare blond und dort schwarz sind! — — Aber die Race und die Sprache ist es nicht allein, was die Völker trennt und verbindet; die Geschichte ist die Hauptperson, welche den verbindenden Kitt mischt, oder das trennende Schwert wetzt.

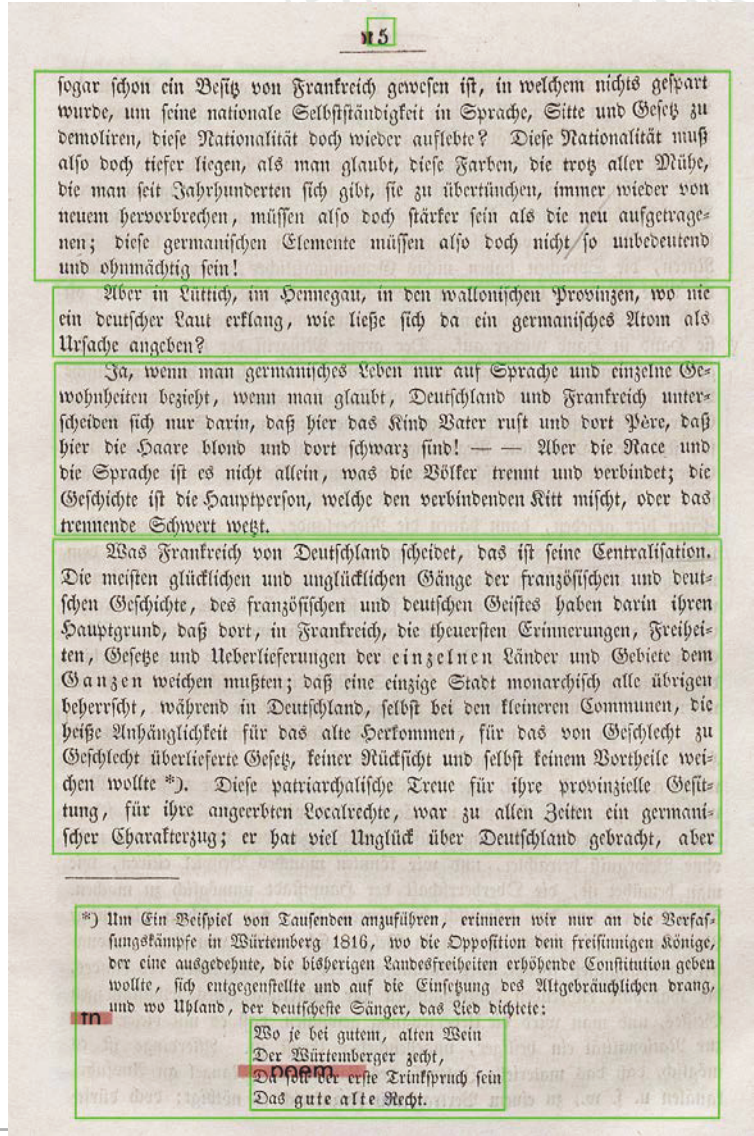
Was Frankreich von Deutschland scheidet, das ist seine Centralisation. Die meisten glücklichen und unglücklichen Gänge der französischen und deutschen Geschichte, des französischen und deutschen Geistes haben darin ihren Hauptgrund, daß dort, in Frankreich, die theuersten Erinnerungen, Freiheiten, Gesetze und Ueberlieferungen der einzelnen Länder und Gebiete dem Ganzen weichen mußten; daß eine einzige Stadt monarchisch alle übrigen beherrscht, während in Deutschland, selbst bei den kleineren Communen, die heiße Anhänglichkeit für das alte Herkommen, für das von Geschlecht zu Geschlecht überlieferte Gesetz, keiner Rücksicht und selbst keinem Vortheile weichen wollte *). Diese patriarchalische Treue für ihre provinzielle Gesittung, für ihre angeerbten Localrechte, war zu allen Zeiten ein germanischer Charakterzug; er hat viel Unglück über Deutschland gebracht, aber

*) Um Ein Beispiel von Tausenden anzuführen, erinnern wir nur an die Verfassungskämpfe in Würtemberg 1816, wo die Opposition dem freisinnigen Könige, der eine ausgedehnte, die bisherigen Landesfreiheiten erhöhende Constitution geben wollte, sich entgegenstellte und auf die Einsetzung des Altgebräuchlichen drang, und wo Uhland, der deutscheste Sänger, das Lied dichtete:

Wo je bei gutem, alten Wein
Der Würtemberger zecht,
Da soll der erste Trinkspruch sein
Das gute alte Recht.

Komponenten eines einfachen OCR-Workflows

1. Bildvorverarbeitung
2. Layoutanalyse
3.



sogar schon ein Besitz von Frankreich gewesen ist, in welchem nichts gespart wurde, um seine nationale Selbstständigkeit in Sprache, Sitte und Gesetz zu demoliren, diese Nationalität doch wieder auflebte? Diese Nationalität muß also doch tiefer liegen, als man glaubt, diese Farben, die trotz aller Mühe, die man seit Jahrhunderten sich gibt, sie zu übertünchen, immer wieder von neuem hervorbrechen, müssen also doch stärker sein als die neu aufgetragenen; diese germanischen Elemente müssen also doch nicht so unbedeutend und ohnmächtig sein!

Aber in Lüttich, im Hennegau, in den wallonischen Provinzen, wo nie ein deutscher Laut erklang, wie ließe sich da ein germanisches Atom als Ursache angeben?

Ja, wenn man germanisches Leben nur auf Sprache und einzelne Gewohnheiten bezieht, wenn man glaubt, Deutschland und Frankreich unterscheiden sich nur darin, daß hier das Kind Vater ruft und dort Père, daß hier die Haare blond und dort schwarz sind! — — Aber die Race und die Sprache ist es nicht allein, was die Völker trennt und verbindet; die Geschichte ist die Hauptperson, welche den verbindenden Kitt mischt, oder das trennende Schwert wetzt.

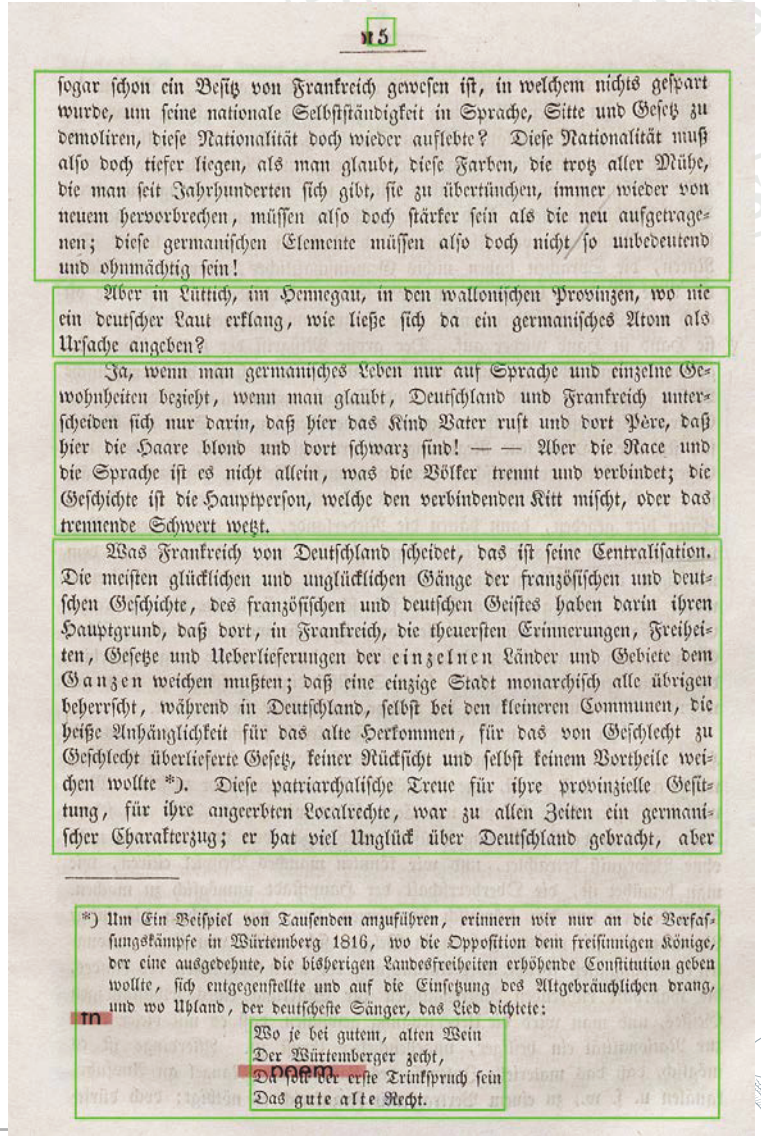
Was Frankreich von Deutschland scheidet, das ist seine Centralisation. Die meisten glücklichen und unglücklichen Gänge der französischen und deutschen Geschichte, des französischen und deutschen Geistes haben darin ihren Hauptgrund, daß dort, in Frankreich, die theuersten Erinnerungen, Freiheiten, Gesetze und Ueberlieferungen der einzelnen Länder und Gebiete dem Ganzen weichen mußten; daß eine einzige Stadt monarchisch alle übrigen beherrscht, während in Deutschland, selbst bei den kleineren Communen, die heiße Anhänglichkeit für das alte Herkommen, für das von Geschlecht zu Geschlecht überlieferte Gesetz, keiner Rücksicht und selbst keinem Vortheile weichen wollte *). Diese patriarchalische Treue für ihre provinzielle Gesittung, für ihre angeerbten Localrechte, war zu allen Zeiten ein germanischer Charakterzug; er hat viel Unglück über Deutschland gebracht, aber

*) Um Ein Beispiel von Tausenden anzuführen, erinnern wir nur an die Verfassungskämpfe in Würtemberg 1816, wo die Opposition dem freisinnigen Könige, der eine ausgedehnte, die bisherigen Landesfreiheiten erhöhende Constitution geben wollte, sich entgegenstellte und auf die Einsetzung des Altgebräuchlichen drang, und wo Uhland, der deutscheste Sänger, das Lied dichtete:

Wo je bei gutem, alten Wein
Der Würtemberger zecht,
Da soll der erste Trinkspruch sein
Das gute alte Recht.

Komponenten eines einfachen OCR-Workflows

1. Bildvorverarbeitung
2. Layoutanalyse
3. Texterkennung



sogar schon ein Besitz von Frankreich gewesen ist, in welchem nichts gespart wurde, um seine nationale Selbstständigkeit in Sprache, Sitte und Gesetz zu demoliren, diese Nationalität doch wieder auflebte? Diese Nationalität muß also doch tiefer liegen, als man glaubt, diese Farben, die trotz aller Mühe, die man seit Jahrhunderten sich gibt, sie zu übertünchen, immer wieder von neuem hervorbrechen, müssen also doch stärker sein als die neu aufgetragenen; diese germanischen Elemente müssen also doch nicht so unbedeutend und ohnmächtig sein!

Aber in Lüttich, im Hennegau, in den wallonischen Provinzen, wo nie ein deutscher Laut erklang, wie ließe sich da ein germanisches Atom als Ursache angeben?

Ja, wenn man germanisches Leben nur auf Sprache und einzelne Gewohnheiten bezieht, wenn man glaubt, Deutschland und Frankreich unterscheiden sich nur darin, daß hier das Kind Vater ruft und dort Père, daß hier die Haare blond und dort schwarz sind! — — Aber die Race und die Sprache ist es nicht allein, was die Völker trennt und verbindet; die Geschichte ist die Hauptperson, welche den verbindenden Kitt mischt, oder das trennende Schwert wetzt.

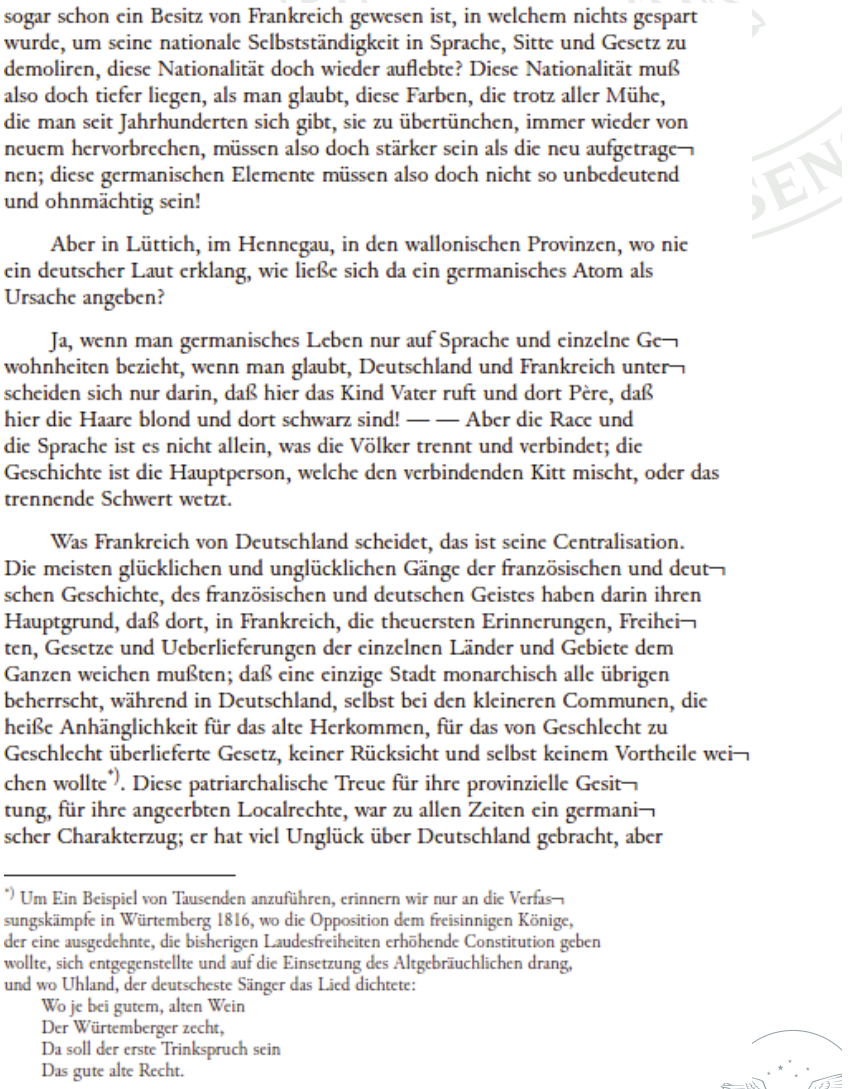
Was Frankreich von Deutschland scheidet, das ist seine Centralisation. Die meisten glücklichen und unglücklichen Gänge der französischen und deutschen Geschichte, des französischen und deutschen Geistes haben darin ihren Hauptgrund, daß dort, in Frankreich, die theuersten Erinnerungen, Freiheiten, Gesetze und Ueberlieferungen der einzelnen Länder und Gebiete dem Ganzen weichen mußten; daß eine einzige Stadt monarchisch alle übrigen beherrscht, während in Deutschland, selbst bei den kleineren Communen, die heiße Anhänglichkeit für das alte Herkommen, für das von Geschlecht zu Geschlecht überlieferte Gesetz, keiner Rücksicht und selbst keinem Vortheile weichen wollte *). Diese patriarchalische Treue für ihre provinzielle Gesittung, für ihre angeerbten Localrechte, war zu allen Zeiten ein germanischer Charakterzug; er hat viel Unglück über Deutschland gebracht, aber

*) Um Ein Beispiel von Tausenden anzuführen, erinnern wir nur an die Verfassungskämpfe in Würtemberg 1816, wo die Opposition dem freisinnigen Könige, der eine ausgedehnte, die bisherigen Landesfreiheiten erhöhende Constitution geben wollte, sich entgegenstellte und auf die Einsetzung des Altgebräuchlichen drang, und wo Uhland, der deutscheste Sänger, das Lied dichtete:

Wo je bei gutem, alten Wein
Der Würtemberger zecht,
Da soll der erste Trinkspruch sein
Das gute alte Recht.

Komponenten eines einfachen OCR-Workflows

1. Bildvorverarbeitung
2. Layoutanalyse
3. Texterkennung



sogar schon ein Besitz von Frankreich gewesen ist, in welchem nichts gespart wurde, um seine nationale Selbstständigkeit in Sprache, Sitte und Gesetz zu demoliren, diese Nationalität doch wieder auflebte? Diese Nationalität muß also doch tiefer liegen, als man glaubt, diese Farben, die trotz aller Mühe, die man seit Jahrhunderten sich gibt, sie zu übertünchen, immer wieder von neuem hervorbrechen, müssen also doch stärker sein als die neu aufgetrage¬ nen; diese germanischen Elemente müssen also doch nicht so unbedeutend und ohnmächtig sein!

Aber in Lüttich, im Hennegau, in den wallonischen Provinzen, wo nie ein deutscher Laut erklang, wie ließe sich da ein germanisches Atom als Ursache angeben?

Ja, wenn man germanisches Leben nur auf Sprache und einzelne Ge¬ wohnheiten bezieht, wenn man glaubt, Deutschland und Frankreich unter¬ scheiden sich nur darin, daß hier das Kind Vater ruft und dort Père, daß hier die Haare blond und dort schwarz sind! — — Aber die Race und die Sprache ist es nicht allein, was die Völker trennt und verbindet; die Geschichte ist die Hauptperson, welche den verbindenden Kitt mischt, oder das trennende Schwert wetzt.

Was Frankreich von Deutschland scheidet, das ist seine Centralisation. Die meisten glücklichen und unglücklichen Gänge der französischen und deut¬ schen Geschichte, des französischen und deutschen Geistes haben darin ihren Hauptgrund, daß dort, in Frankreich, die theuersten Erinnerungen, Freihei¬ ten, Gesetze und Ueberlieferungen der einzelnen Länder und Gebiete dem Ganzen weichen mußten; daß eine einzige Stadt monarchisch alle übrigen beherrscht, während in Deutschland, selbst bei den kleineren Communen, die heiße Anhänglichkeit für das alte Herkommen, für das von Geschlecht zu Geschlecht überlieferte Gesetz, keiner Rücksicht und selbst keinem Vortheile wei¬ chen wollte*). Diese patriarchalische Treue für ihre provinzielle Gesit¬ tung, für ihre angeerbten Localrechte, war zu allen Zeiten ein germani¬ scher Charakterzug; er hat viel Unglück über Deutschland gebracht, aber

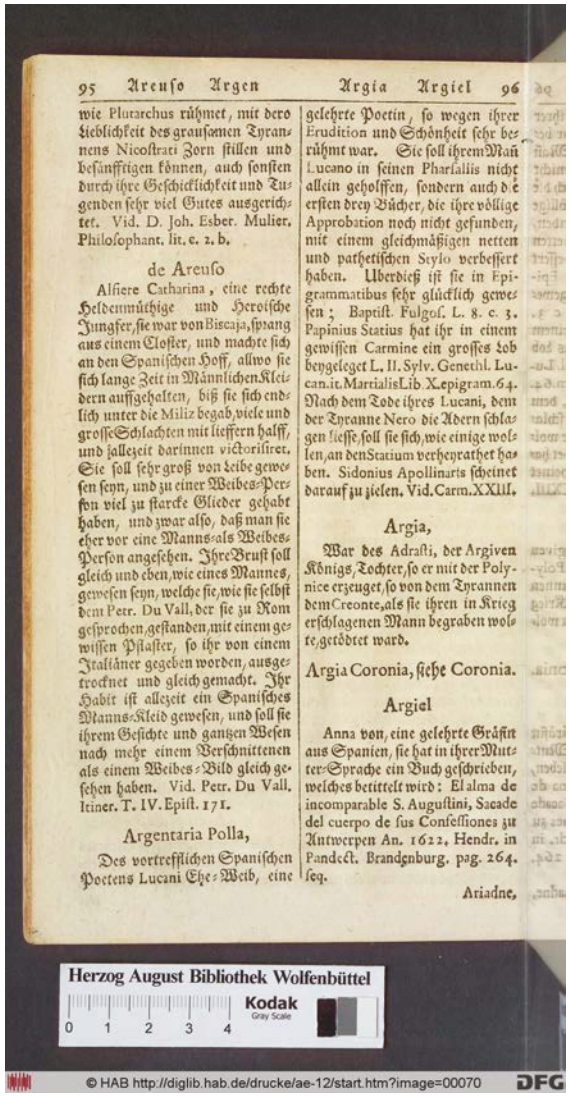
*) Um Ein Beispiel von Tausenden anzuführen, erinnern wir nur an die Verfas¬ sungskämpfe in Würtemberg 1816, wo die Opposition dem freisinnigen Könige, der eine ausgedehnte, die bisherigen Laudesfreiheiten erhöhende Constitution geben wollte, sich entgegenstellte und auf die Einsetzung des Altgebräuchlichen drang, und wo Uhland, der deutscheste Sänger das Lied dichtete:

Wo je bei gutem, alten Wein
Der Würtemberger zecht,
Da soll der erste Trinkspruch sein
Das gute alte Recht.

Komponenten eines OCR-Workflows: Bildvorverarbeitung

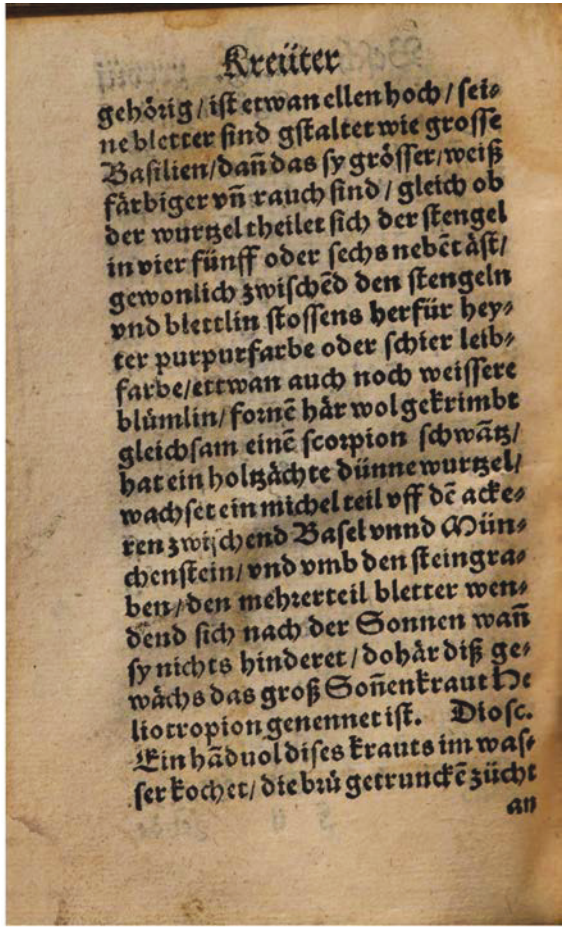
- Prozesse zur bestmöglichen Vorbereitung der Digitalisate für OLR und OCR
  - **Cropping:** Beschneidung des Digitalisats auf den Druckbereich
  - **Deskewing:** Rotation des Digitalisats zur Begradigung von Schrägstellungen
  - **Binarization:** Binäre Kodierung der Pixel (bedruckte Bereiche schwarz, 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

Komponenten eines einfachen OCR-Workflows: Cropping



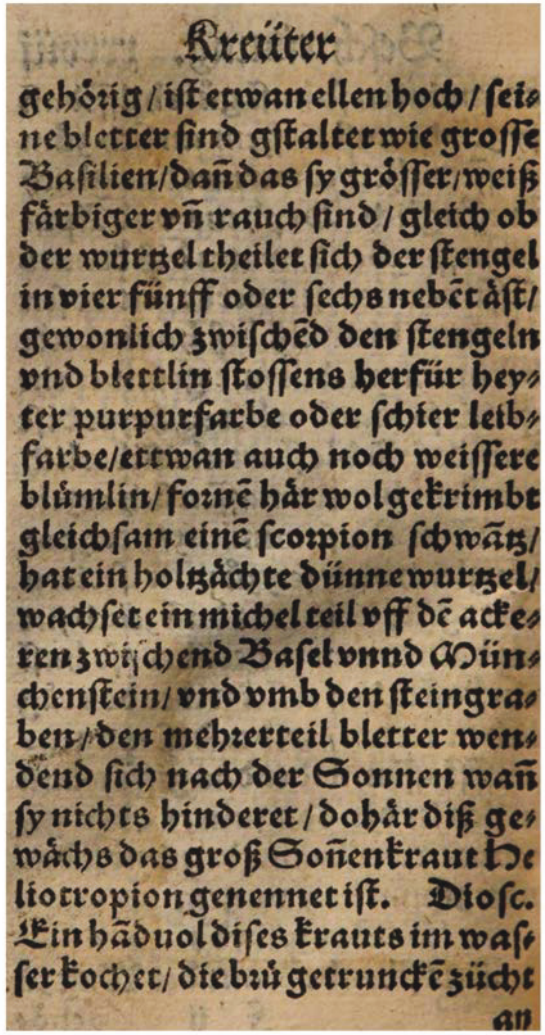
95 Areuso Argen Argia Argiel 96
wie Plutarchus rühmet, mit dero Lieblichkeit des grausamen Tyrannens Nicostrati Zorn stillen und besänfftigen können, auch sonsten durch ihre Geschicklichkeit und Tugenden sehr viel Gutes ausgerichtet. Vid. D. Joh. Esber. Mulier. Philosophant. lit. e. 2. b.
de Areuso
Alfiere Catharina, eine rechte Heldenmüthige und Heroische Jungfer, sie war von Biscaja, sprang aus einem Closter, und machte sich an den Spanischen Hoff, allwo sie sich lange Zeit in Männlichen Kleidern auffgehalten, biß sie sich endlich unter die Miliz begab, viele und grosse Schlachten mit liefern halff, und allezeit darinnen victorisiret. Sie soll sehr groß von Leibe gewesen seyn, und zu einer Weibes-Person viel zu starcke Glieder gehabt haben, und zwar also, daß man sie eher vor eine Manns- als Weibes-Person angesehen. Ihre Brust soll gleich und eben, wie eines Mannes, gewesen seyn, welche sie, wie sie selbst dem Petr. Du Vall, der sie zu Rom gesprochen, gestanden, mit einem gewissen Pflaster, so ihr von einem Italiäner gegeben worden, ausgetrocknet und gleich gemacht. Ihr Habit ist allezeit ein Spanisches Manns-Kleid gewesen, und soll sie ihrem Gesichte und gantzen Wesen nach mehr einem Verschnittenen als einem Weibes-Bild gleich gesehen haben. Vid. Petr. Du Vall. Itiner. T. IV. Epist. 171.
Argentaria Polla,
Des vortrefflichen Spanischen Poetens Lucani Ehe-Weib, eine gelehrte Poetin, so wegen ihrer Erudition und Schönheit sehr berühmt war. Sie soll ihrem Mann Lucano in seinen Pharsaliis nicht allein geholffen, sondern auch die ersten drey Bücher, die ihre völlige Approbation noch nicht gefunden, mit einem gleichmäßigen netten und pathetischen Stylo verbessert haben. Überdieß ist sie in Epigrammatibus sehr glücklich gewesen; Baptist. Fulgos. L. 8. c. 3. Papinius Statius hat ihr in einem gewissen Carmine ein grosses Lob beygeleget L. II. Sylv. Genethl. Lucan. it. Martialis Lib. X. epigram. 64. Nach dem Tode ihres Lucani, dem der Tyranne Nero die Adern schlagen liesse, soll sie sich, wie einige wollen, an den Statium verheyrathet haben. Sidonius Apollinaris scheinet darauf zu zielen. Vid. Carm. XXIII.
Argia,
War des Adrasti, der Argiven Königs, Tochter, so er mit der Polynice erzeuget, so von dem Tyrannen dem Creonte, als sie ihren in Krieg erschlagenen Mann begraben wolte, getödtet ward.
Argia Coronia, siehe Coronia.
Argiel
Anna von, eine gelehrte Gräfin aus Spanien, sie hat in ihrer Mutter-Sprache ein Buch geschrieben, welches betittelt wird: El alma de incomparable S. Augustini, Sacade del cuerpo de sus Confessiones zu Antwerpen An. 1622. Hendr. in Pandect. Brandenburg. pag. 264. seq.
Ariadne,

Komponenten eines einfachen OCR-Workflows: Deskewing



Kreüter

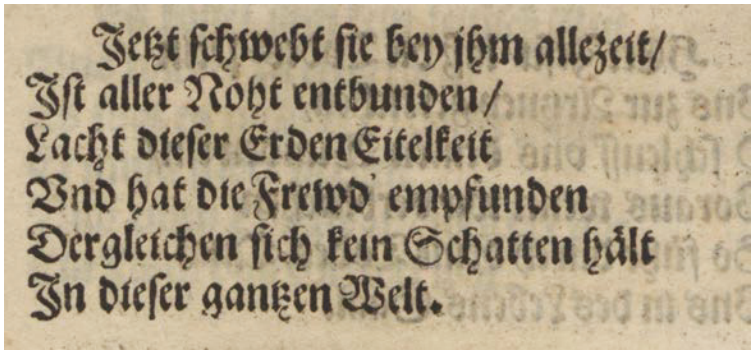
gehörig / ist etwan ellen hoch / sei-
ne bletter sind gstaltet wie grosse
Basilien / dañ das sy grösser / weiß
färbiger vñ rauch sind / gleich ob
der wurtzel theilet sich der stengel
in vier fünff oder sechs nebēt äst /
gewonlich zwischēd den stengeln
vnd blettlin stossens herfür hey-
ter purpurfarbe oder schier leib-
farbe / ettwan auch noch weissere
blümlin / fornē här wol gekrimbt
gleichsam einē scorpion schwātz /
hat ein holtzächte dünne wurtzel /
wachset ein michel teil vff dē acke-
ren zwischend Basel vnnd Mün-
chenstein / vnd vmb den steingra-
ben / den mehrerteil bletter wen-
dend sich nach der Sonnen wañ
sy nichts hinderet / dohär diß ge-
wächs das groß Soñenkraut He
liotropion genennet ist. Diosc.
Ein hāduol dises krauts im was-
ser kochet / die brü getrunckē zücht
an



Kreüter

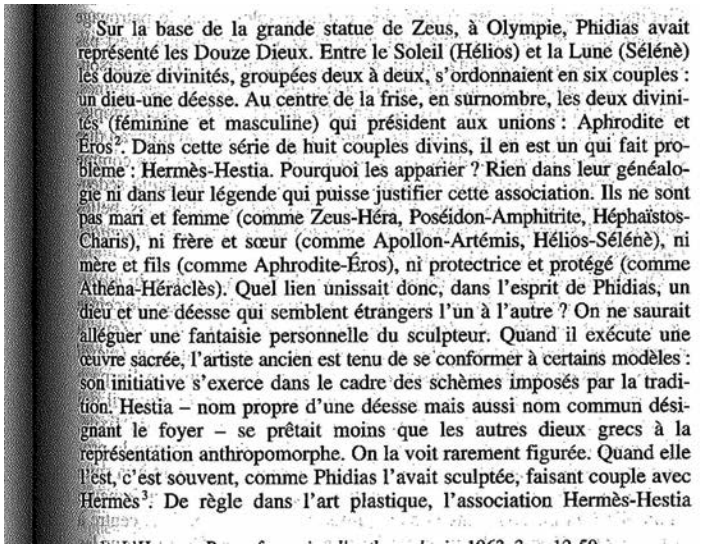
gehörig / ist etwan ellen hoch / sei-
ne bletter sind gstaltet wie grosse
Basilien / dañ das sy grösser / weiß
färbiger vñ rauch sind / gleich ob
der wurtzel theilet sich der stengel
in vier fünff oder sechs nebēt äst /
gewonlich zwischēd den stengeln
vnd blettlin stossens herfür hey-
ter purpurfarbe oder schier leib-
farbe / ettwan auch noch weissere
blümlin / fornē här wol gekrimbt
gleichsam einē scorpion schwātz /
hat ein holtzächte dünne wurtzel /
wachset ein michel teil vff dē acke-
ren zwischend Basel vnnd Mün-
chenstein / vnd vmb den steingra-
ben / den mehrerteil bletter wen-
dend sich nach der Sonnen wañ
sy nichts hinderet / dohär diß ge-
wächs das groß Soñenkraut He
liotropion genennet ist. Diosc.
Ein hāduol dises krauts im was-
ser kochet / die brü getrunckē zücht
an

Komponenten eines OCR-Workflows: Binarization

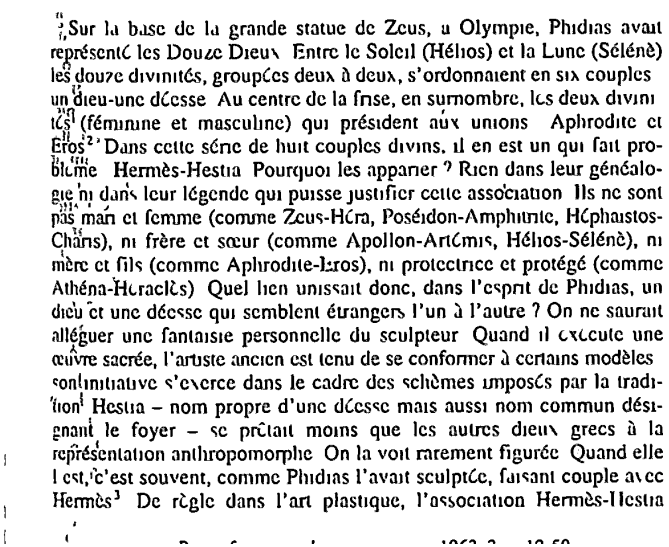


Jetzt ſchwebt ſie bey jhm allezeit/
Iſt aller Nöht entbunden/
Lacht dieſer Erden Eitelkeit
Vnd hat die Frewd' empfunden
Dergleichen ſich kein Schatten hält
In dieſer gantzen Welt.

Komponenten eines OCR-Workflows: Despeckling



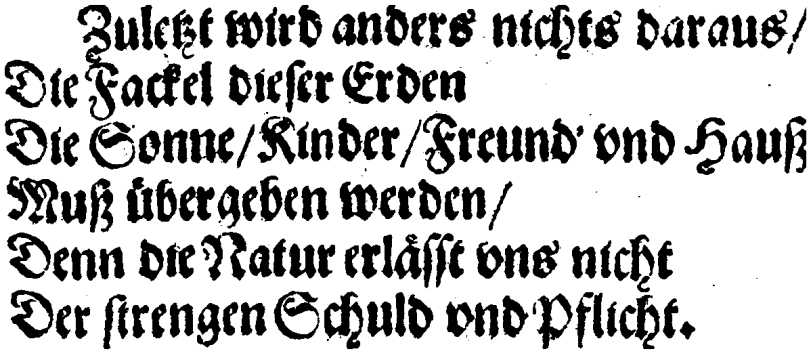
Sur la base de la grande statue de Zeus, à Olympie, Phidias avait représenté les Douze Dieux. Entre le Soleil (Hélios) et la Lune (Séléné) les douze divinités, groupées deux à deux, s'ordonnaient en six couples : un dieu-une déesse. Au centre de la frise, en surnombre, les deux divinités (féminine et masculine) qui président aux unions : Aphrodite et Éros[2]. Dans cette série de huit couples divins, il en est un qui fait problème : Hermès-Hestia. Pourquoi les apparier ? Rien dans leur généalogie ni dans leur légende qui puisse justifier cette association. Ils ne sont pas mari et femme (comme Zeus-Héra, Poséidon-Amphitrite, Héphaïstos-Charis), ni frère et sœur (comme Apollon-Artémis, Hélios-Séléné), ni mère et fils (comme Aphrodite-Éros), ni protectrice et protégé (comme Athéna-Héraclès). Quel lien unissait donc, dans l'esprit de Phidias, un dieu et une déesse qui semblent étrangers l'un à l'autre ? On ne saurait alléguer une fantaisie personnelle du sculpteur. Quand il exécute une œuvre sacrée, l'artiste ancien est tenu de se conformer à certains modèles : son initiative s'exerce dans le cadre des schèmes imposés par la tradition. Hestia – nom propre d'une déesse mais aussi nom commun désignant le foyer – se prêtait moins que les autres dieux grecs à la représentation anthropomorphe. On la voit rarement figurée. Quand elle l'est, c'est souvent, comme Phidias l'avait sculptée, faisant couple avec Hermès[3]. De règle dans l'art plastique, l'association Hermès-Hestia

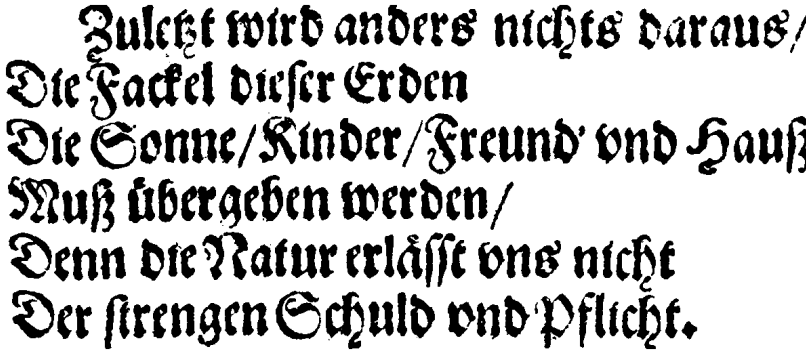


Sur la base de la grande statue de Zeus, à Olympie, Phidias avait représenté les Douze Dieux Entre le Soleil (Hélios) et la Lune (Séléné) les douze divinités, groupées deux à deux, s'ordonnaient en six couples un dieu-une déesse Au centre de la frise, en surnombre, les deux divinités (féminine et masculine) qui président aux unions Aphrodite et Éros[2] Dans cette série de huit couples divins, il en est un qui fait problème Hermès-Hestia Pourquoi les apparier ? Rien dans leur généalogie ni dans leur légende qui puisse justifier cette association Ils ne sont pas mari et femme (comme Zeus-Héra, Poséidon-Amphitrite, Héphaïstos-Charis), ni frère et sœur (comme Apollon-Artémis, Hélios-Séléné), ni mère et fils (comme Aphrodite-Éros), ni protectrice et protégé (comme Athéna-Héraclès) Quel lien unissait donc, dans l'esprit de Phidias, un dieu et une déesse qui semblent étrangers l'un à l'autre ? On ne saurait alléguer une fantaisie personnelle du sculpteur Quand il exécute une œuvre sacrée, l'artiste ancien est tenu de se conformer à certains modèles son initiative s'exerce dans le cadre des schèmes imposés par la tradition Hestia – nom propre d'une déesse mais aussi nom commun désignant le foyer – se prêtait moins que les autres dieux grecs à la représentation anthropomorphe On la voit rarement figurée Quand elle l'est, c'est souvent, comme Phidias l'avait sculptée, faisant couple avec Hermès[3] De règle dans l'art plastique, l'association Hermès-Hestia

Komponenten eines einfachen OCR-Workflows: Dewarping

Komponenten eines OCR-Workflows: Einfluss Binarisierung


Zuletzt wird anders nichts daraus/
Die Fackel dieſer Erden
Die Sonne/Kinder/Freund' vnd Hauß
Muß übergeben werden/
Denn die Natur erläſſt vns nicht
Der ſtrengen Schuld vnd Pflicht.


Zuletzt wird anders nichts daraus/
Die Fackel dieſer Erden
Die Sonne/Kinder/Freund' vnd Hauß
Muß übergeben werden/
Denn die Natur erläſſt vns nicht
Der ſtrengen Schuld vnd Pflicht.

Zuletzt wird anders nichts darans/
Dir zacke1 dieſer Erden r ' '
Die Sonne/Kindrr/Frenud' vnd Hauß
Muß übergeben werden/ ' '
Denn dirNatnr erläſſt vns' mehr '
Der ſtreugenSchnld ondPflichr.

Zuletzt wird anders nichts darans/
Die Fackel dieſer Erden
Die Sonne/Kinder/Frennnd' vnd Hauß
Muß übergeben werden/
Denn deeNainr erläſſt vns nicht
Der ſtrengen Schuld vndPflicht.

Komponenten eines OCR-Workflows: Bildvorverarbeitung

Werkzeuge:

- Bestandteil der meisten OCR-Programme, häufig jedoch nicht modular
- spezielle Tools
  - **Scantailor** `https://github.com/scantailor/scantailor`
    + umfassendes, frei verfügbares Werkzeug
    - keine Programmierschnittstelle (API)
  - **Olena/SCRIBO**
    `https://www.lrde.epita.fr/wiki/Olena/Modules#SCRIBO`
    + frei verfügbare Programmierbibliothek für Deskewing, Binarisierung (Implementierung verschiedener Ansätze)
    - keine Weiterentwicklung/Pflege, schlechtes API-Design
  - **Unpaper** `https://github.com/Flameeyes/unpaper`
    + frei verfügbare Programmierbibliothek für Deskewing und Despeckling

Komponenten eines OCR-Workflows: Bildvorverarbeitung

Werkzeuge:

- teilweise auch in Bildbearbeitungsbibliotheken integriert
  - **ImageMagick** `https://www.imagemagick.org/`
    + extrem umfangreiches, frei verfügbares Softwarepaket
    - keine spezifische OCR-Implementierung (aber: `http://www.fmwconcepts.com/imagemagick/`)
  - **Leptonica** `http://www.leptonica.com/`
    + sehr umfangreiches, frei verfügbares Softwarepaket
    + Anwendung in `Tesseract`
- zahlreiche **wissenschaftliche Veröffentlichungen** zu einzelnen Aspekten
- **wissenschaftliche Wettbewerbe** zu ausgewählten Aspekten (insb. Binarization und Deskewing)
- Forschungsergebnisse finden **kaum Eingang in die Praxis**

Komponenten eines OCR-Workflows: Layoutanalyse

- 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

Komponenten eines OCR-Workflows: Layoutanalyse

- **strukturierende** Elemente
  - Absätze
  - Überschriften

Anfangsgründe der physischen Geographie. 13

sind langgestreckt, und ihre Reihe ist entweder mit der Küste gleichlaufend (Nordseeinseln), oder sie erscheinen als Fortsetzung der Bergketten des benachbarten Festlandes (die Cykladen an den Küsten Griechenlands). Die oceanischen Inseln, einzeln in der Regel von rundlichem Umriß, sind meistens zu Gruppen vereinigt (Sandwichinseln). — Von besonders eigenthümlicher Entstehungsart und Form sind die Korallen-inseln (vgl. Fig. 5 und 6).

<abb> <abb>

Fig. 5. Insel White-Sunday (Niedr. Inseln). Fig. 6. Insel Bola Bola (Gesellschaftsinseln).

Gewäßer des Festlandes.

§. 11. **Regen und Quellen.** (G. §. 13.) Das Meer ist die Mutter aller Gewäßer des Festlandes. Unter der Einwirkung der Sonnenwärme erheben sich, besonders in den wärmeren Gegenden, große Massen süßen Waßers als Dampf bis zu bedeutender Höhe in die Atmosphäre, werden hier zu Wolken verdichtet, die der Wind fortführt, bis sie sich soweit verdichtet haben, daß das Waßer, in flüßiger oder fester Form als Regen oder Schnee zur Erde herabstürzt. Die Menge des auf solche Weise der Erdoberfläche zurückgegebenen Waßers wird auf täglich 5 Kubikmeilen geschätzt. Ein Theil dieses Niederschlags dringt in die Erdoberfläche ein, sammelt sich auf seinem unterirdischen Wege zu Waßeradern und tritt zuletzt wieder in Gestalt von Quellen zu Tage. Die verschiedenen Quellen zeigen wesentliche Unterschiede hinsichtlich des Wärmegrades, wie in Bezug auf die Stoffe, die ihr Waßer aus dem Innern der Erde mitbringt. (Warme Quellen, Salz- und Schwefelquellen, Sauerbrunnen u. s. w.)

§. 12. **Flüße und Seen.** (G. §. 14.) Das Waßer der hervorbrechenden Quellen sucht nun, durch die Schwerkraft geleitet, die jedesmal tiefste Stelle des Bodens auf und erreicht, wenn es nicht im Glutsande der Wüste verdunstet, zuletzt wieder das Meer oder sammelt sich in großen

Komponenten eines OCR-Workflows: Layoutanalyse

- **strukturierende** Elemente
  - Absätze
  - Überschriften
- **textflussunterbrechende** Elemente
  - Seitenzahlen
  - Kolumnentitel
  - Abbildungsunterschriften
  - Marginalien etc.

Anfangsgründe der physischen Geographie. 13

sind langgestreckt, und ihre Reihe ist entweder mit der Küste gleichlaufend (Nordseeinseln), oder sie erscheinen als Fortsetzung der Bergketten des benachbarten Festlandes (die Cykladen an den Küsten Griechenlands). Die oceanischen Inseln, einzeln in der Regel von rundlichem Umriß, sind meistens zu Gruppen vereinigt (Sandwichinseln). — Von besonders eigenthümlicher Entstehungsart und Form sind die Korallen-inseln (vgl. Fig. 5 und 6).

<abb> <abb>

Fig. 5. Insel White-Sunday (Niedr. Inseln). Fig. 6. Insel Bola Bola (Gesellschaftsinseln).

4. Gewäßer des Festlandes.

§. 11. **Regen und Quellen.** (G. §. 13.) Das Meer ist die Mutter aller Gewäßer des Festlandes. Unter der Einwirkung der Sonnenwärme erheben sich, besonders in den wärmeren Gegenden, große Massen süßen Waßers als Dampf bis zu bedeutender Höhe in die Atmosphäre, werden hier zu Wolken verdichtet, die der Wind fortführt, bis sie sich soweit verdichtet haben, daß das Waßer, in flüßiger oder fester Form als Regen oder Schnee zur Erde herabstürzt. Die Menge des auf solche Weise der Erdoberfläche zurückgegebenen Waßers wird auf täglich 5 Kubikmeilen geschätzt. Ein Theil dieses Niederschlags dringt in die Erdoberfläche ein, sammelt sich auf seinem unterirdischen Wege zu Waßeradern und tritt zuletzt wieder in Gestalt von Quellen zu Tage. Die verschiedenen Quellen zeigen wesentliche Unterschiede hinsichtlich des Wärmegrades, wie in Bezug auf die Stoffe, die ihr Waßer aus dem Innern der Erde mitbringt. (Warme Quellen, Salz- und Schwefelquellen, Sauerbrunnen u. s. w.)

§. 12. **Flüße und Seen.** (G. §. 14.) Das Waßer der hervorbrechenden Quellen sucht nun, durch die Schwerkraft geleitet, die jedesmal tiefste Stelle des Bodens auf und erreicht, wenn es nicht im Glutsande der Wüste verdunstet, zuletzt wieder das Meer oder sammelt sich in großen

Komponenten eines OCR-Workflows: Layoutanalyse

- **strukturierende** Elemente
  - Absätze
  - Überschriften
- **textflussunterbrechende** Elemente
  - Seitenzahlen
  - Kolumnentitel
  - Abbildungsunterschriften
  - Marginalien etc.
- **nichttextuelle** Elemente
  - Abbildungen
  - Tabellen

Anfangsgründe der physischen Geographie. 13

sind langgestreckt, und ihre Reihe ist entweder mit der Küste gleichlaufend (Nordseeinseln), oder sie erscheinen als Fortsetzung der Bergketten des benachbarten Festlandes (die Cykladen an den Küsten Griechenlands). Die oceanischen Inseln, einzeln in der Regel von rundlichem Umriß, sind meistens zu Gruppen vereinigt (Sandwichinseln). — Von besonders eigenthümlicher Entstehungsart und Form sind die Korallen-inseln (vgl. Fig. 5 und 6).

<abb> <abb>

Fig. 5. Insel White-Sunday (Niedr. Inseln). Fig. 6. Insel Bola Bola (Gesellschaftsinseln).

Gewäßer des Festlandes.

§. 11. **Regen und Quellen.** (G. §. 13.) Das Meer ist die Mutter aller Gewäßer des Festlandes. Unter der Einwirkung der Sonnenwärme erheben sich, besonders in den wärmeren Gegenden, große Massen süßen Waßers als Dampf bis zu bedeutender Höhe in die Atmosphäre, werden hier zu Wolken verdichtet, die der Wind fortführt, bis sie sich soweit verdichtet haben, daß das Waßer, in flüßiger oder fester Form als Regen oder Schnee zur Erde herabstürzt. Die Menge des auf solche Weise der Erdoberfläche zurückgegebenen Waßers wird auf täglich 5 Kubikmeilen geschätzt. Ein Theil dieses Niederschlags dringt in die Erdoberfläche ein, sammelt sich auf seinem unterirdischen Wege zu Waßeradern und tritt zuletzt wieder in Gestalt von Quellen zu Tage. Die verschiedenen Quellen zeigen wesentliche Unterschiede hinsichtlich des Wärmegrades, wie in Bezug auf die Stoffe, die ihr Waßer aus dem Innern der Erde mitbringt. (Warme Quellen, Salz- und Schwefelquellen, Sauerbrunnen u. s. w.)

§. 12. **Flüße und Seen.** (G. §. 14.) Das Waßer der hervorbrechenden Quellen sucht nun, durch die Schwerkraft geleitet, die jedesmal tiefste Stelle des Bodens auf und erreicht, wenn es nicht im Glutsande der Wüste verdunstet, zuletzt wieder das Meer oder sammelt sich in großen

Komponenten eines OCR-Workflows: Layoutanalyse

Werkzeuge:

- auch bei der OLR **Missverhältnis** zwischen Forschungsergebnissen und verfügbaren Lösungen
- OCR-Programme implementieren einfache Lösungen zur Page Segmentation, teilweise separat adressierbar
  - Klassifizierung beschränkt sich im Wesentlichen auf Text vs. Nichttext
  - Qualität auf schwierigen Vorlagen überschaubar
- wissenschaftliche Wettbewerbe und Untersuchungen befassen sich mit der Erkennung **komplexer Layouts** und **Dokumentstukturierung**
- jedoch **keine umfassenden** Ansätzen (Kustoden, Marginalien, Bogensignaturen)
- ebenfalls kaum Forschung zu **polygonen Zonen**

Komponenten eines OCR-Workflows: Layoutanalyse

Werkzeuge:

- einzelner Befehl für Seiten- und Zeilensegmentierung in `OCRopus`
  - im Ergebnis nur Einzelbilder auf Zeilenebene
  - **keine Koordinaten**, kein Zugriff auf Seitensegmentierung
- Zugriff auf alle Ebenen der Seitensegmentierung in `Tesseract`
  - **inklusive Koordinaten**
  - basale Klassifizierung der Segmente (Spalten, Abbildungen, Formeln, Tabellen, Text)
- Layouterkennungswerkzeug `Larex` *(Reul at al. 2017)*
  - Festlegung buchspezifischer Parameter durch den Nutzer (Spalten, Kolumnentitel etc.) für Klassifizierungsaufgabe
  - manuelle Nachkorrektur über Benutzeroberfläche
  - inklusive Zeilenerkennung, open-source: `https://github.com/chreul/LAREX`, keine API

Komponenten eines OCR-Workflows: Texterkennung

- 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)*

Komponenten eines OCR-Workflows: Texterkennung

Werkzeuge:

- überraschend viele verfügbare OCR-Engines
- `https://github.com/kba/awesome-ocr`
- Abbyy `FineReader` am verbreitetsten im produktiven Einsatz
- zwei Platzhirsche im **Open-Source-Bereich**
- `Tesseract` `https://github.com/tesseract-ocr/tesseract`
  - ursprünglich von Hewlett-Packard entwickelt
  - von Google übernommen und Open-Source gestellt
  - viele **mitgelieferte Modelle** (auch für Fraktur)
  - mit Version 4 Umstieg auf zeilenorientierte Erkennung auf Basis neuronaler Netze

Komponenten eines OCR-Workflows: Texterkennung

Werkzeuge:

- `OCRopus` `https://github.com/tmbdev/ocropy`
  - entwickelt von Thomas Breul mit Unterstützung von Google
  - ursprünglich als Wrapper für Tesseract, später mit eigener Erkennungsroutine auf Basis neuronaler Netze
  - nur wenige **mitgelieferte Modelle**
- `Gamera` `https://github.com/hsnr-gamera/gamera`
  - komplettes Framework für Layoutanalyse und Texterkennung
  - zeichenorientierter Ansatz auf Basis des „k nearest neighbor“-Algorithmus'
  - nur ein **mitgeliefertes Modell**

Modelltraining

Modelltraining

- Texterkennung auf Basis **statistischer** Modelle
  - Induktion anhand manuell erstellter Trainingsdaten (i.e. **Ground Truth**)
  - Wahrscheinlichkeitsverteilung abhängig vom Modelltyp entweder direkt berechnet oder (iterativ) optimiert
- unterschiedliche Ansätze erfordern unterschiedliche **Trainingsprozeduren**
- grundsätzliches Vorgehen jedoch gleich → **Alignierung** von Text und Bild
  - unterschiedliche Anforderung an **Annotationstiefe**
  - Qualität und Quantität der Trainingsdaten bestimmt Qualität der Modelle
- Kompromiss zwischen **Übertragbarkeit** und spezifischer Textqualität
  - mitgelieferte Modelle häufig zu „allgemein“
  - Qualität der Texterkennung im Vergleich zu Standardmodellen **signifikant höher** *(Springmann et al. 2015)*

Modelltraining: Trainingsdaten

- Digitalisate und zugehöriger, **fehlerfreier** Volltext
- Alignierung auf Zeichen- oder Zeilenebene
- **zeichenorientierte** Ansätze: jedes Zeichen mindestens einmal im Trainingsmaterial
- **zeilenorientierte** Ansätze: ca. 10 Seiten eines Buches
- `Tesseracts` „Latin model“ (i.e. großmaßstäbliches Mehrsprachenmodell für Antiquaschriftarten): ca. 400 000 Zeilen in ca. 4 500 Schriftarten

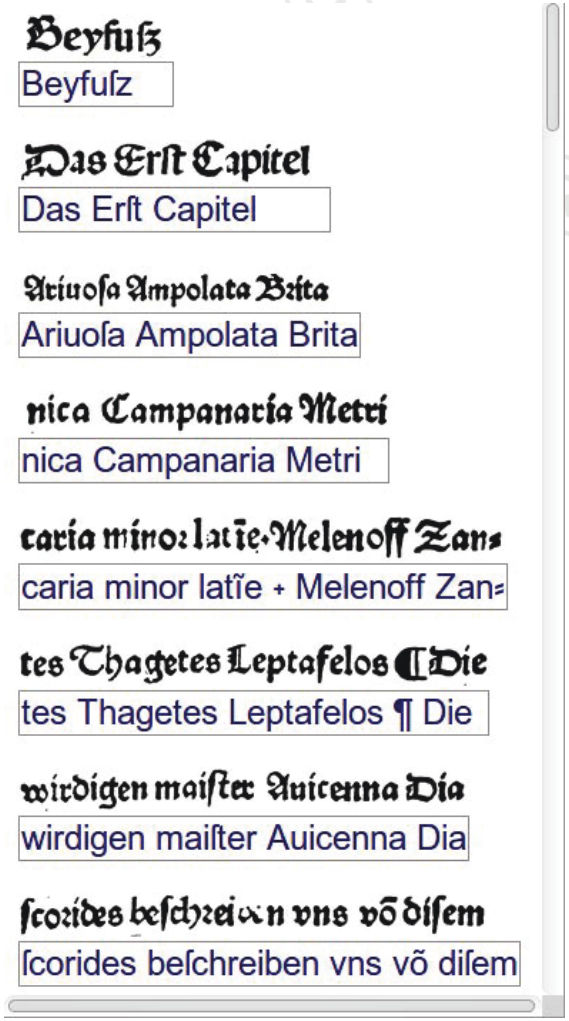
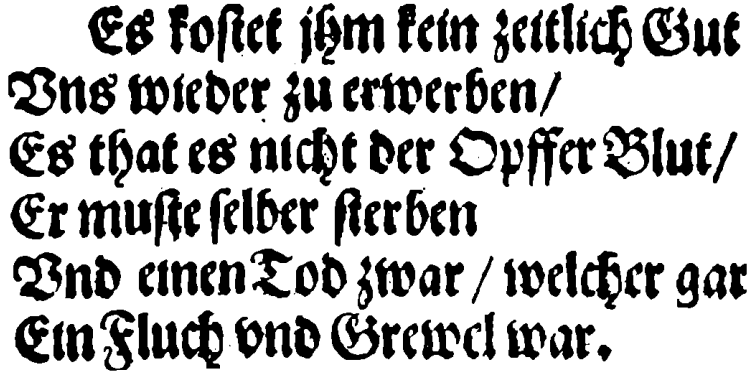


Image by Uwe Springmann, CC BY-SA 4.0

Modelltraining: Trainingseffekte


Es koſtet jhm kein zeitlich Gut
Vns wieder zu erwerben/
Es that es nicht der Opffer Blut/
Er muſte ſelber ſterben
Vnd einen Tod zwar / welcher gar
Ein Fluch vnd Grewel war.

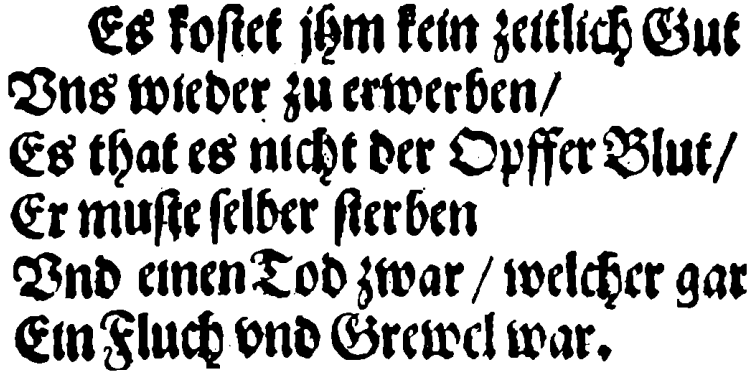
Abbyy `Finereader 11`

ES kostet Om kein zeitlich Gut
Dns wieder zu erwerben/
ES that es nicht der OpfferBluk/
Cr muste selber sterben
Vnd emenTod zwar/ welcher gar
EinFluch vnd Grcwcl war.

`OCRopus`

Es kostet jzm kein zattlchGut
Bns wteder zu crwerben?
Es that cs mcht der OpfferBlut?
Ermustcselbcr stcrbcn
Bnd emnenTodzwar, welchae gar
EtFluchvnd Grcwewar.

Modelltraining: Trainingseffekte


Es koſtet jhm kein zeitlich Gut
Vns wieder zu erwerben/
Es that es nicht der Opffer Blut/
Er muſte ſelber ſterben
Vnd einen Tod zwar / welcher gar
Ein Fluch vnd Grewel war.

Abbyy `Finereader 11`

ES kostet Om kein zeitlich Gut
Dns wieder zu erwerben/
ES that es nicht der OpfferBluk/
Cr muste selber sterben
Vnd emenTod zwar/ welcher gar
EinFluch vnd Grcwcl war.

`OCRopus` (homebrew)

Es koſlet jhm kein zeitlich Gut
Vns wieder zu terwerben/
Es that es nicht der Opffer Blut/
Er muſte ſelber ſterben
Vnd einen Tod zwar/ welcher gar
Ein Fluch vnd Grewel war.

Modelltraining: Werkzeuge

- jede OCR-Software kommt mit **eigener Trainingsprozedur**
- zahlreiche „ease-of-use“-Wrapper:
  - `Tesseract`: `VietOCR`, `Aletheia`, `Franken++`
  - `OCRopus`: `OCRocis`, HTML-Wrapper
- Probleme:
  - teilweise **kostenpflichtig** (`Aletheia`) bzw. nicht mehr gepflegt (`Franken++`)
  - teilweise zu **umständlich** in der Bedienung (`OCRopus`-Training, `Aletheia`)

Optimierungsoptionen

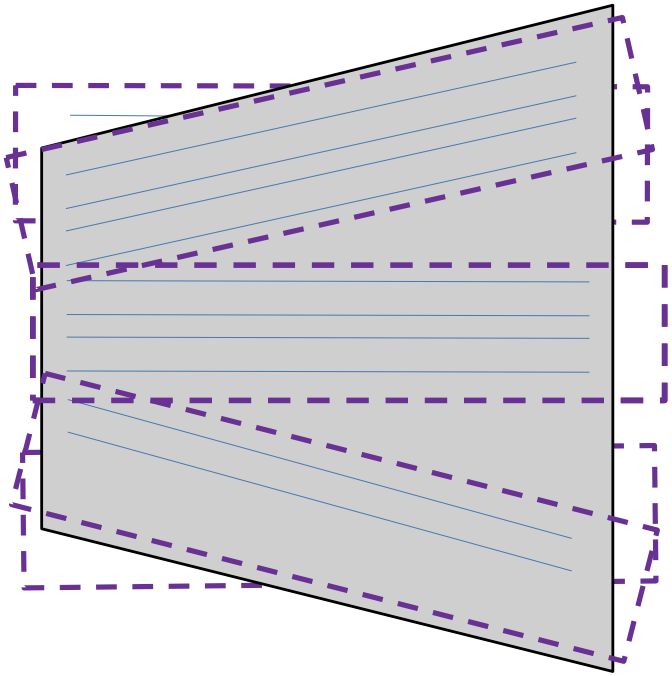
Optimierungsoptionen: Lokale Bildoptimierung

- historische Vorlagen bzw. ältere Digitalisate oftmals suboptimal für OCR
  - unterschiedliche **Beleuchtung**
  - charakteristische **Trapezform**
- verschiedene Bearbeitungsebenen
  - Dokument, Seite, Absatz (bzw. Textzone), Zeile
  - Operationen greifen **wiederholt** auf verschiedenen Ebenen ein
  - **maximale Adaptivität** bzgl. spezifischer Charakteristika auf Bild- und Textebene
  - Rekonstruierbarkeit über Metadaten (z.B. Koordinaten) zu gewährleisten

Optimierungsoptionen: Lokale Bildoptimierung

Rezept:

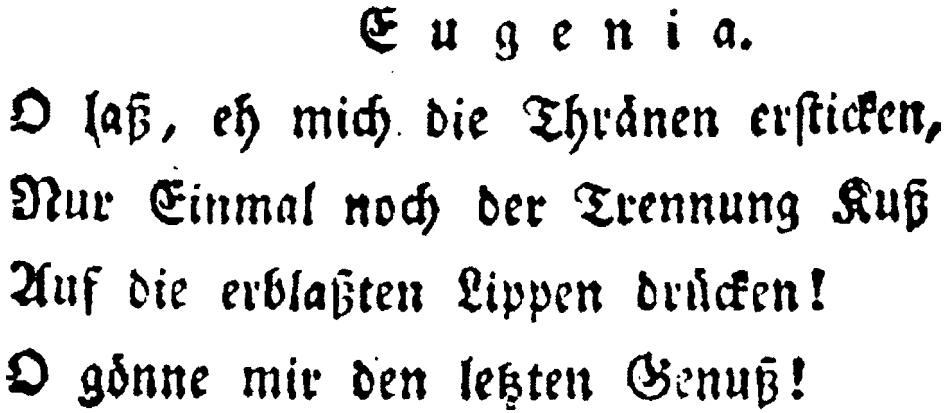
1. **Bildoptimierung** auf Seitenebene
2. **Seitensegmentierung** auf Seitenebene
3. **Extraktion** der Segmente aus dem (nicht-optimierten) Original
4. **Bildoptimierung** auf Segmentebene
5. **Zeilensegmentierung** auf Segmentebene
6. **Extraktion** der Zeile aus dem (nichtoptimierten) Segment
7. **Bildoptimierung** auf Zeilenebene



Optimierungsoptionen: OCR-Merging

- 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)*

OCR-Merging: Kotzebue „Schutzgeist“ (1814)



Eugenia.

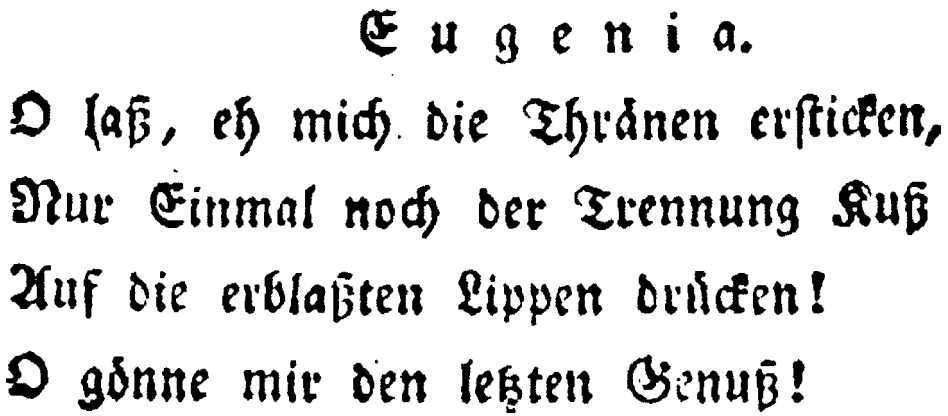
O laß, eh mich die Thränen erſticken,
Nur Einmal noch der Trennung Kuß
Auf die erblaßten Lippen drücken!
O gönne mir den letzten Genuß!

Abbyy Finereader

Eugenia.

O §aß, eh mich die Thränen ersticken.
Nur Einmal noch der Trennung Kuß
Auf die erblaßten Lippen drücken!
O gönne mir den lehten Genuß!

OCR-Merging: Kotzebue „Schutzgeist“ (1814)



Eugenia.

O laß, eh mich die Thränen ersticken,
Nur Einmal noch der Trennung Kuß
Auf die erblaßten Lippen drücken!
O gönne mir den letzten Genuß!

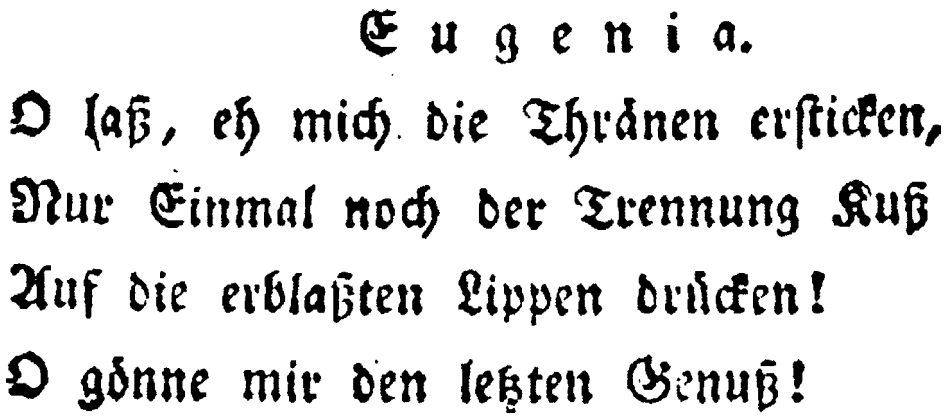
Tesseract

Eugeaia.
O laß, H) mich. die Tht*äaea erlHckm,
Nur Einqu noch der Trennung Kuß
Auf die erqußteu Lippen drückenx
O gönne mie den legten Genußx

Abbyy Finereader

Eugenia.
O §aß, eh mich die Thränen ersticken.
Nur Einmal noch der Trennung Kuß
Auf die erblaßten Lippen drücken!
O gönne mir den lehten Genuß!

OCR-Merging: Kotzebue „Schutzgeist“ (1814)



Eugenia.

O laß, eh mich die Thränen ersticken,
Nur Einmal noch der Trennung Kuß
Auf die erblaßten Lippen drücken!
O gönne mir den letzten Genuß!

Abbyy Finereader

Eugenia.
O §aß, eh mich die Thränen ersticken.
Nur Einmal noch der Trennung Kuß
Auf die erblaßten Lippen drücken!
O gönne mir den lehten Genuß!

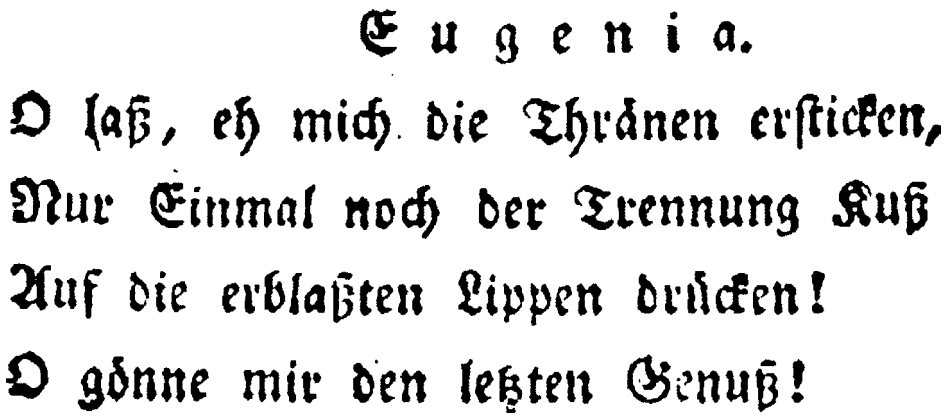
Tesseract

Eugeaia.
O laß, H) mich. die Tht*äaea erlHckm,
Nur Einqu noch der Trennung Kuß
Auf die erqußteu Lippen drückenx
O gönne mie den legten Genußx

OCRopus

Eugenia.
O haß, eh mich die Thränen ersticken,
Nur Einmal noch der Trennung Kuß
Auf die erblaßten Lippen drücken!
O gdnne mir den letzten Genuß!

OCR-Merging: Kotzebue „Schutzgeist“ (1814)



E u g e n i a.

O laß, eh mich die Thränen erſticken,
Nur Einmal noch der Trennung Kuß
Auf die erblaßten Lippen drücken!
O gönne mir den letzten Genuß!

Merge

E u g e n i a.
O laß, eh mich die Thränen ersticken,
Nur Einmal noch der Trennung Kuß
Auf die erblaßten Lippen drücken!
O gönne mir den letzten Genuß!

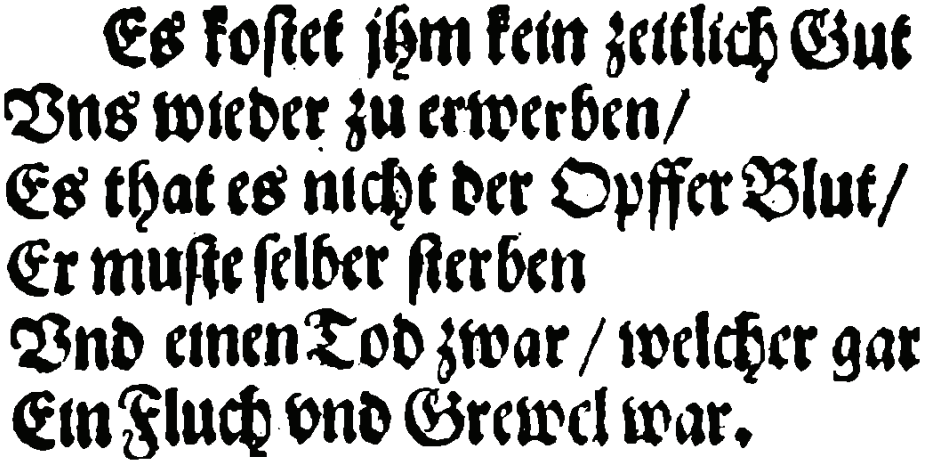
Tesseract

Eugeaia.
O laß, H) mich. die Tht*äaea erlHckm,
Nur Einqu noch der Trennung Kuß
Auf die erqußteu Lippen drückenx
O gönne mie den legten Genußx

OCRopus

E u g e n i a.
O **h**aß, eh mich die Thränen ersticken,
Nur Einmal noch der Trennung Kuß
Auf die erblaßten Lippen drücken!
O g**d**nne mir den letzten Genuß!

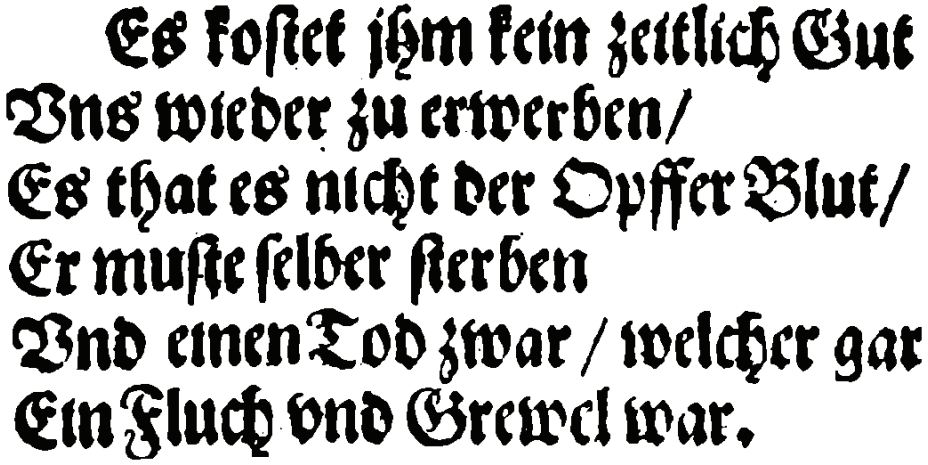
OCR-Merging: Dach „Einfältige Leichreime“ (1653)



Abbyy Finereader

ES kostet Om kein zeitlich Gut
Dns wieder zu erwerben/
ES that es nicht der OpfferBluk/
Cr muste selber sterben
Vnd emenTod zwar/ welcher gar
EinFluch vnd Grcwcl war.

OCR-Merging: Dach „Einfältige Leichreime“ (1653)


Es koſtet jhm kein zeitlich Gut
Vns wieder zu erwerben/
Es that es nicht der Opffer Blut/
Er muſte ſelber ſterben
Vnd einen Tod zwar / welcher gar
Ein Fluch vnd Grewel war.

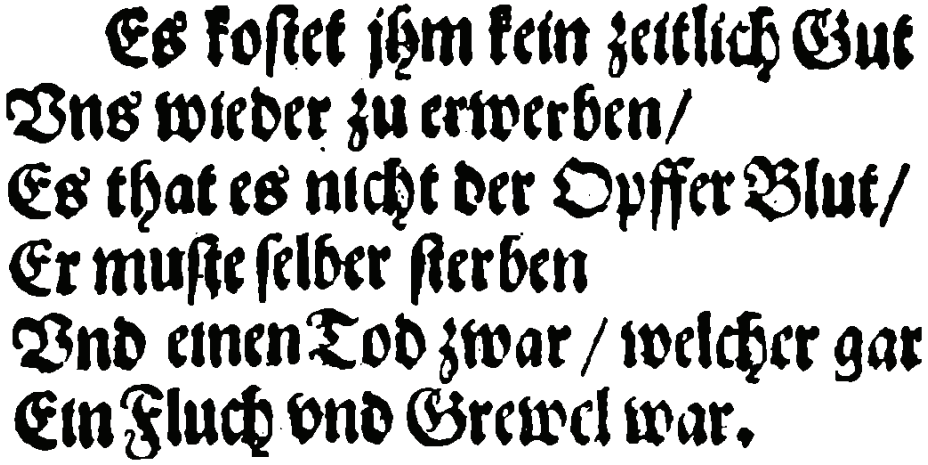
Abbyy Finereader

ES kostet Om kein zeitlich Gut
Dns wieder zu erwerben/
ES that es nicht der OpfferBluk/
Cr muste selber sterben
Vnd emenTod zwar/ welcher gar
EinFluch vnd Grcwcl war.

Tesseract

Es koſtet jhm kein zeitlich Gut
Vns wieder zu erwerben/
Es ihaietz?i1ichi der Opffer Blui/
Er muſte ſelber ſterben
Vnd einenTod zwar / welcher gar
EinFliich'vud Grewel war.

OCR-Merging: Dach „Einfältige Leichreime“ (1653)


Es koſtet jhm kein zeitlich Gut
Vns wieder zu erwerben/
Es that es nicht der Opffer Blut/
Er muſte ſelber ſterben
Vnd einen Tod zwar / welcher gar
Ein Fluch vnd Grewel war.

Abbyy Finereader

ES kostet Om kein zeitlich Gut
Dns wieder zu erwerben/
ES that es nicht der OpfferBluk/
Cr muste selber sterben
Vnd emenTod zwar/ welcher gar
EinFluch vnd Grcwcl war.

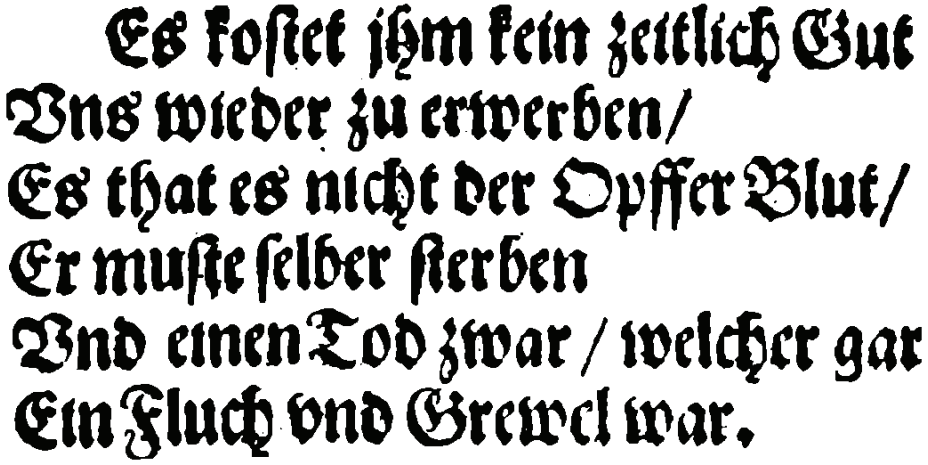
Tesseract

Es koſtet jhm kein zeitlich Gut
Vns wieder zu erwerben/
Es ihaietz?i1ichi der Opffer Blui/
Er muſte ſelber ſterben
Vnd einenTod zwar / welcher gar
EinFliich'vud Grewel war.

OCRopus

Es koſlet jhm kein zeitlich Gut
Vns wieder zu terwerben/
Es that es nicht der Opffer Blut/
Er muſte ſelber ſterben
Vnd einen Tod zwar/ welcher gar
Ein Fluch vnd Grewel war.

OCR-Merging: Dach „Einfältige Leichreime“ (1653)


Es koſtet jhm kein zeitlich Gut
Vns wieder zu erwerben/
Es that es nicht der Opffer Blut/
Er muſte ſelber ſterben
Vnd einen Tod zwar / welcher gar
Ein Fluch vnd Grewel war.

Merge

Es koſtet jhm kein zeitlich Gut
Vns wieder zu erwerben/
Es that es nicht der Opffer Blut/
Er muſte ſelber ſterben
Vnd einen Tod zwar / welcher gar
Ein Fluch vnd Grewel war.

Tesseract

Es koſtet jhm kein zeitlich Gut
Vns wieder zu erwerben/
Es ihaietz?i1ichi der Opffer Blui/
Er muſte ſelber ſterben
Vnd einenTod zwar / welcher gar
EinFliich'vud Grewel war.

OCRopus

Es koſlet jhm kein zeitlich Gut
Vns wieder zu terwerben/
Es that es nicht der Opffer Blut/
Er muſte ſelber ſterben
Vnd einen Tod zwar/ welcher gar
Ein Fluch vnd Grewel war.

Optimierungsoptionen: OCR-Nachkorrektur

- auch unter optimierten Bedingungen verbleiben OCR-Fehler
- manuelle oder automatische Korrektur des Textes zur Erhöhung der Qualität
- drei Ansatzmöglichkeiten:
  - **manuell** (Collaborative Manual Correction/ Crowdsourcing)
  - **programmunterstützt** (Interactive Postcorrection)
  - **automatisch**
- „klassische“ Aufgabe der **Computerlinguistik**
  - Anleihen bei Rechtschreibkorrektur
  - bzw. Schreibungsnormalisierung *(Jurish 2012)*

Optimierungsoptionen: OCR-Nachkorrektur

Werkzeuge

- **manuell:**
  - manuelle Transkription/Korrektur des OCR-Ergebnisses, erfordert umfassende Konzeption und (anfängliche) Betreuung, bietet Ansatz für Gamification
  - diverse proprietäre und Open-Source-Lösungen, **plattformgebunden**, z.B. `DTAQ`
- **programmunterstützt:**
  - Unterstützung der manuellen Korrektur durch **Korrekturvorschläge** und Hervorhebung wahrscheinlich fehlerhafter Texterkennungsergebnisse
  - **Post Correction Tool** `https://github.com/cisocrgroup/PoCoTo`
- **automatisch:**
  - Korrektur auf Basis von (lexikalischen) Ground-Truth-Daten
  - **Rechtschreibkorrekturprogramme** wie `hunspell` `http://hunspell.github.io/`
  - projektspezifische (Insel)-Lösungen wie der sog. **Bremer Ansatz** für die Zeitschrift „Die Grenzboten“ *(Nölte et al. 2016)*

Optimierungsoptionen: Synthetisches Trainingsmaterial

- Volltexte historischer Drucke zunehmend vorhanden
  - manuelle Erfassung normalerweise **ohne** Text-Bild-Alignierung
  - Erstellung von Trainingsmaterial **zeitaufwendig** und teuer
- Idee: Einsatz von **Font-rendering**-Software um automatisch alignierte Trainingsdaten zu erzeugen
  - Verwendung historischer Schriften, z.B. Fraktur
    `http://www.ligafaktur.de/Schriften.html`
  - „künstliche“ Artefakte zur Nachahmung der Druckalterung

Optimierungsoptionen: Synthetisches Trainingsmaterial

Werkzeuge:

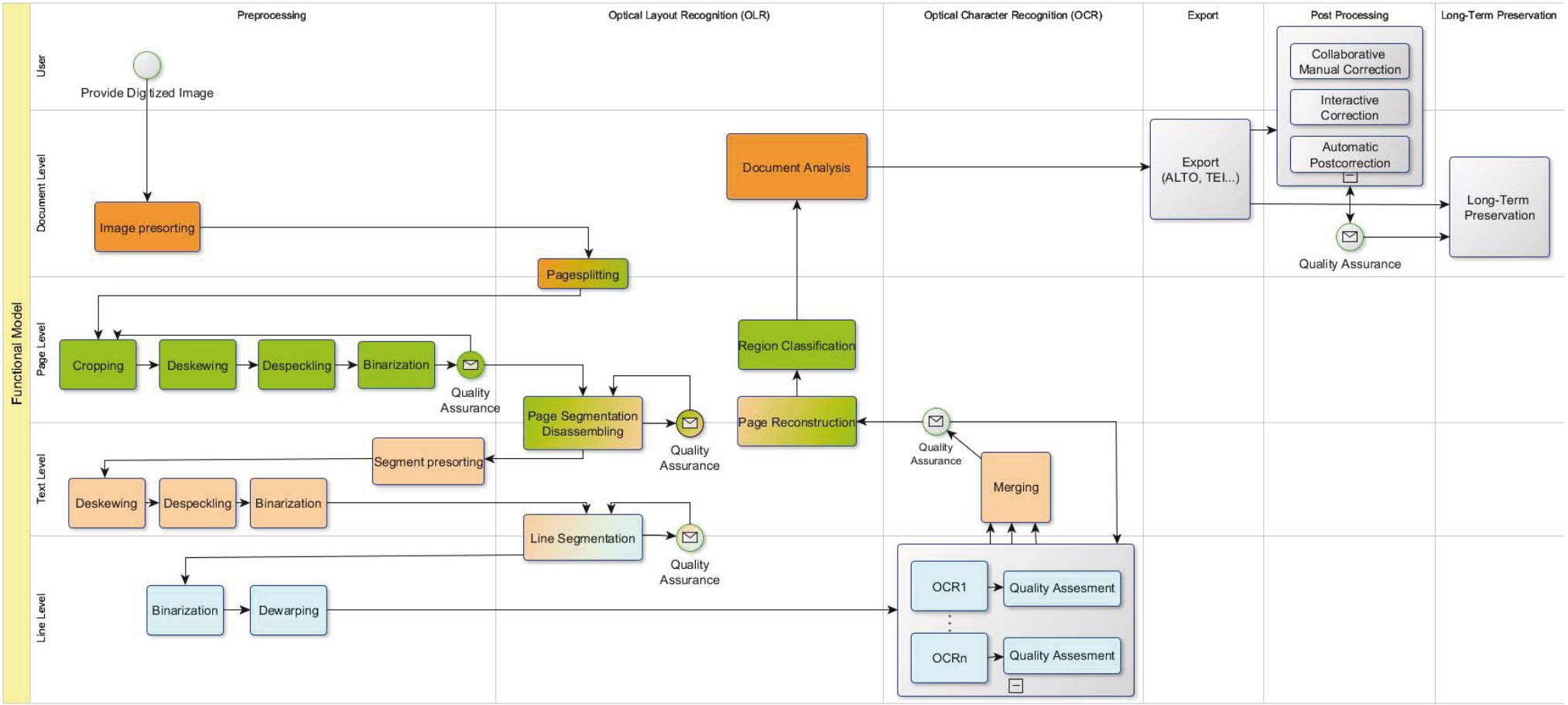
- `OCRopus` (und sein Fork `Kraken`) und `Tesseract` mit **Generierungsmechanismus**
- viele Projekte zur Erstellung **historischer Fonts** im TTF/OTF-Format für (praktisch) alle alphabetischen Schriftsysteme

Komplexere OCR-Workflows

Komplexere OCR-Workflows

- „einfache“ OCR-Workflows in allen OCR-Lösungen implementiert
- **keine Möglichkeit** zur direkten Integration der diskutierten Optimierungsmöglichkeiten
- kein modulares **Workflowmanagmentsystem** im Bereich OCR vorhanden
- momentane Lösung:
  - Zugriff auf **einzelne Module**
  - Kombination in **spezifischem Workflow**
  - aka. Skripte und Hacks
- **OCR-D**

Komplexere OCR-Workflows: Beispiel



OCR-D

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

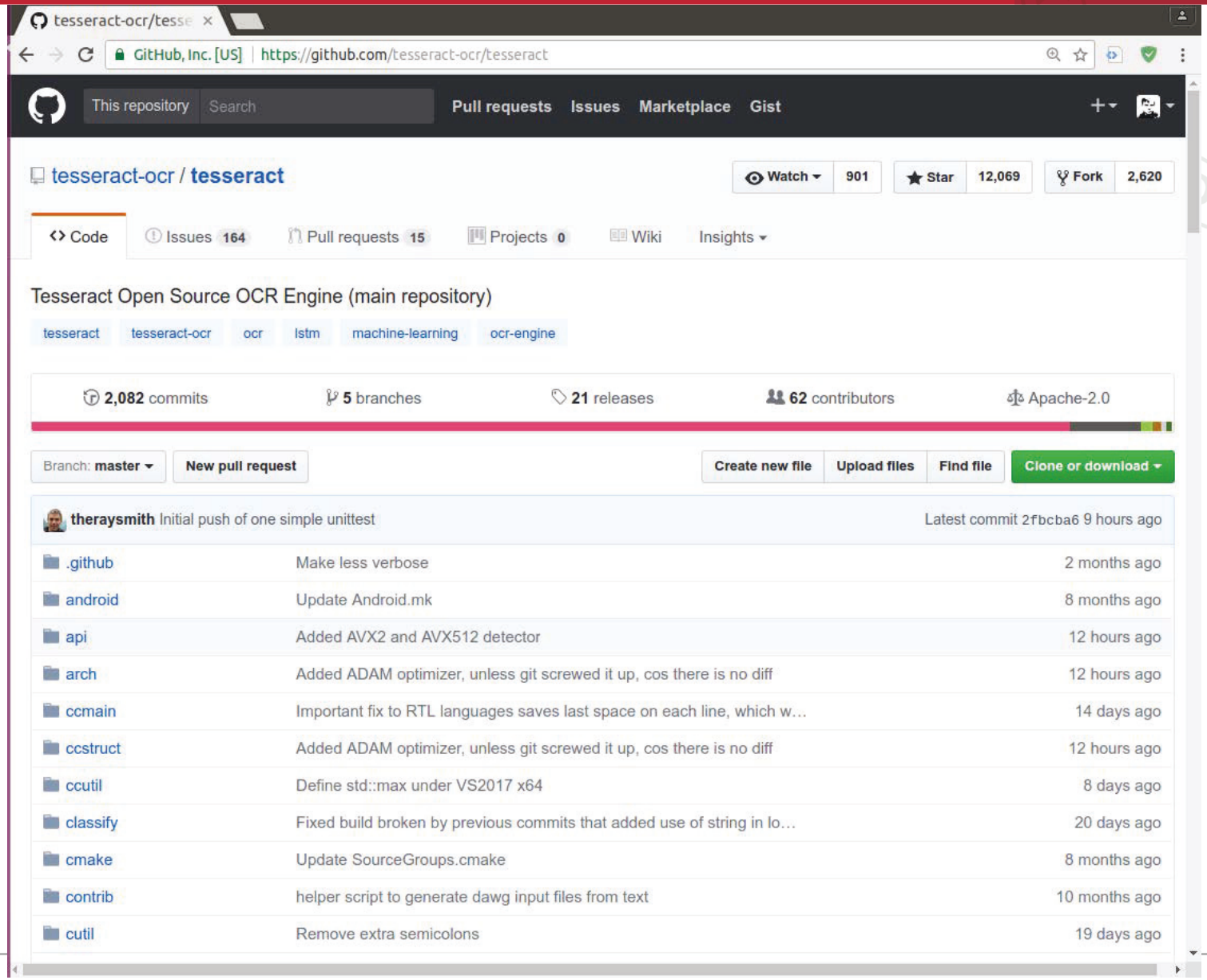
OCR-D: Projektprämissen

- **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

OCR-D: Open-Source-Paradigma

- öffentlich geförderte Projekte $\rightsquigarrow$ öffentlich verfügbare Projektergebnisse
- dank **Digital Humanities** „Kulturrevolution“
  - Daten (i.e. Texte) veröffentlicht unter CC
  - „Belohnung“ durch wissenschaftliche Veröffentlichung und Zitierungen
  - **reproducible science**
- im Bereich der Werkzeuge: **Verbesserungsbedarf**
  - Entwicklung von Tools mit hohem Aufwand und öffentlich gefördert
  - am Projektende Veröffentlichung eines Archivs mit Quellcode unter Open-Source-Lizenz
- Ziel: Einbindung der Nutzercommunity von **Anfang an**
  - Fehlermeldung und Funktionalitätsfeedback während der Entwicklung
  - Weiterentwicklung und Pflege auch nach Ablauf der Förderung

OCR-D: Open-Source-Paradigma



Danke für Ihre Aufmerksamkeit!

OCR-D-Team: Elisa Hermann, Maria Federbusch, Clemens Neudecker, Ajinkya Prabhune, **Matthias Boenig**

Mehr zu OCR: `https://www.zotero.org/groups/ocr-d`