



(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?

- Optical Character Recognition: Automatische Erfassung von Text in Bildern
- ursprünglich begrenzt auf Zeichenerkennung
- heute häufig Synonym für den gesamten Texterfassungsprozess
 - ▶ Bildvorverarbeitung
 - ▶ Layoutanalyse (OLR)
 - ▶ Zeilenerkennung
 - ▶ ...



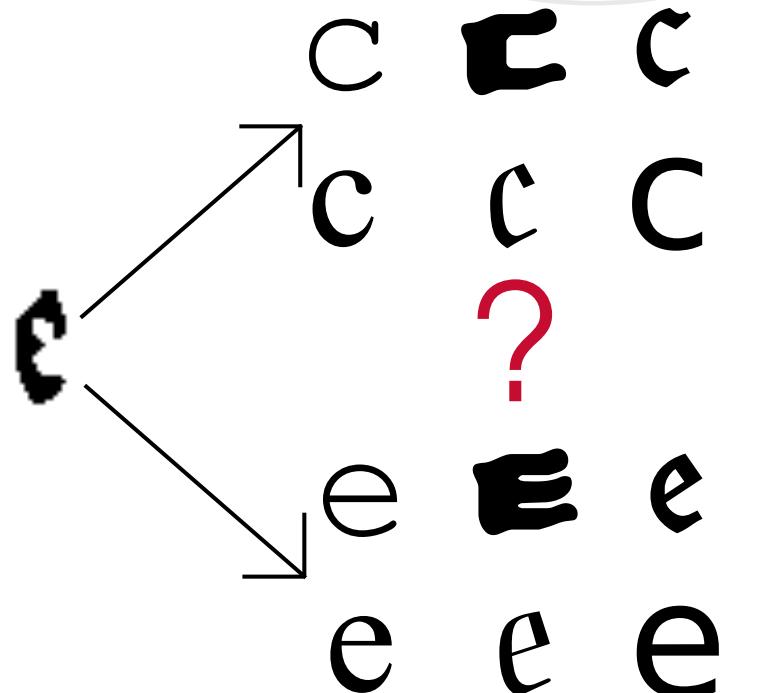
Zeichenorientierte Ansätze

- 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

Umfangsgründe der physischen Geographie.

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

(Breuel 2008)





Wozu braucht man OCR?

Wozu braucht man OCR?

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



4YCH428



4YCH428



4	Y	C	H	4	2	8
---	---	---	---	---	---	---

Wozu braucht man OCR?

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



Wozu braucht man OCR?

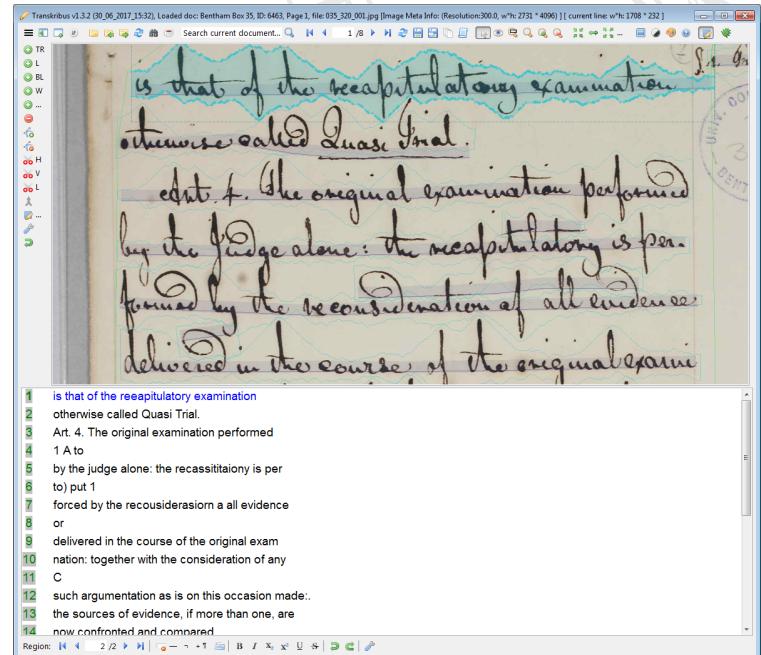
■ typische Anwendungen:

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



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 sūmū bonū et beatitudinē abs-
qz om̄i deliberatōne vel p̄electōne Vnde dicit au-
gustinus in soliloquīs. Deus quē amat omne qđ
amare potest: siue sciens: siue nesciens. Circa neu-
trā istarū est meritū vel demeritū: quia nec volū-
tas. virtus em̄ & viciū voluntaria sunt. Volun-
taria aut̄ diuidic̄ in duas: scilicet amiciciā & con-
cupiscentiā. Amicicia diligim̄ illud quod ppter
se diligimus. Concupiscētia vero diligimus illud
cui bonū volum̄: scz ad delectandū in eo. vtro-
qz istorū modorū diligimus deū naturalit̄: & ange-
li etiā in primo statu. Sed diligebat angelus deū
sup om̄ia amore occupiscentie. scz in ip̄o delectan-
do sup om̄ia. Nec tñ seq̄tur q̄ haberet caritatem
quia nō diligebat deū ppter ip̄m deū sed ppter se :



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 fūmū bonū et beatitudinē abf · q3 om̄i deliberatōne vel p̄electōne Vnde dicit au guftinus in foliloquijis · Deus quē amat omne qd amare potest:fiue sciens:fiue nefsciens· Circa neu trā iftarū eft meritū vel demeritū : quia nec volū tas · virtus em̄ & vitiū voluntaria funt · Volun
taria aūt diuidl̄ in duas: fcilicet amiciciā & con
cupiscentiā · Amicicia diligim⁹ illud quod ppter
se diligimus · Concupiscētia vero diligimus illud
cui bonū volum⁹ : fc̄ ad delectandū in eo · vtro
q3 ifto₂ modo₂ diligimus deū naturalit̄: & ange
li etiā in primo statu · Sed diligebat angelus deū
sup om̄ia amore ccupiscentie · f3 in iþo delectan
do sup om̄ia · Nec tñ feqtur q̄ haberet caritatem
quia nō diligebat deū ppter iþm deū sed ppter fe :



Warum überhaupt OCR?

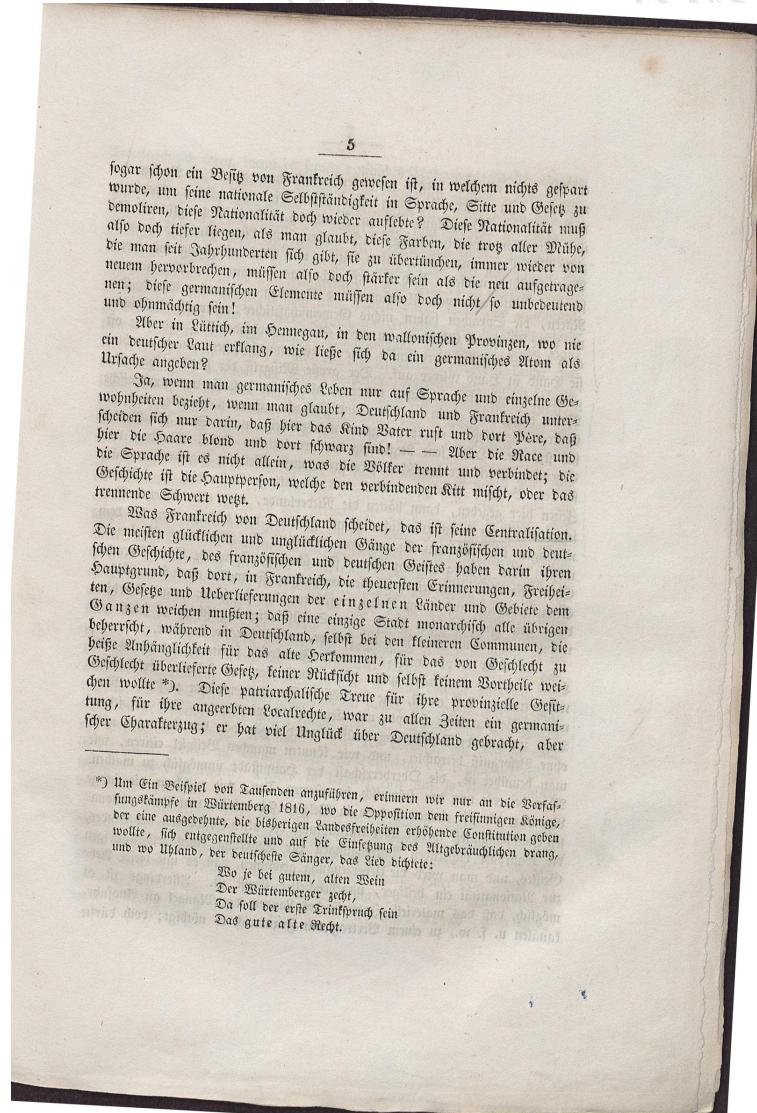
- OCR ist immer **fehlerhaft**! Aber:
- verändertes „Rechercheverhalten“ in Zeiten zunehmender Verfügbarkeit digitaler Quellen
 - ▶ Wissenserwerb durch Internetsuche
 - ▶ Sekundärliteratur (fast) vollständig **textdigital** verfügbar
 - ▶ Navigationssystem vs. Autoatlas
- Ansprüche an Verfügbarkeit von **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



sogar schon ein Besitz von Frankreich gewesen ist, in welchem nichts gespart wurde, um seine nationale Selbstständigkeit in Sprache, Sitte und Gesetz zu demonstrieren, diese Nationalität doch wieder aufzubauen? Diese Nationalität muss also doch tiefer liegen, als man glaubt, diese Karben, die trotz aller Mühe, die man seit Jahrhunderten sich gibt, sie zu überwinden, 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 erslang, 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 unterscheiden sich nur darin, daß hier das Kind Batur rast und dort Pore, daß hier die Haare blond und dort schwarz sind! — — Aber die Rasse und die Sprache ist es nicht allein, was die Völker trennt und verbindet; die Geschichte ist die Hauptperson, welche den verbindenden Leit mischt, über das trennende Schwert wegt.

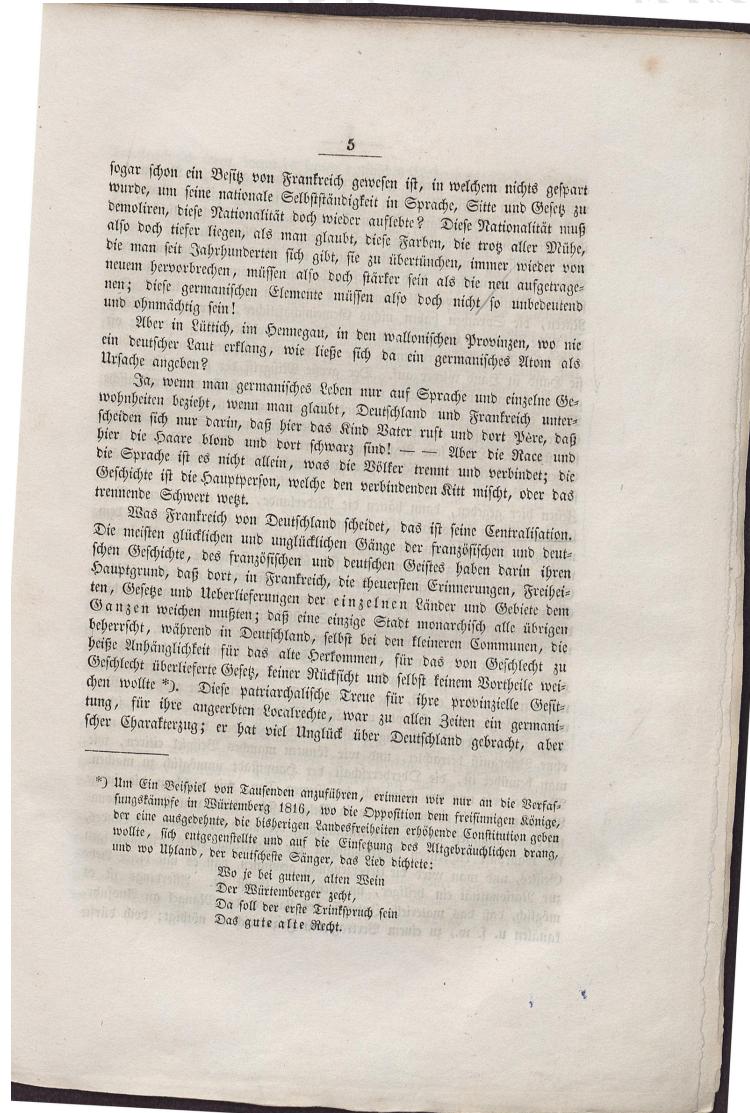
Was Frankreich von Deutschland scheidet, das ist seine Centralisation. Die meisten glücklichen und unglücklichen Gänge der französischen und deutschen Geschichte, des französischen und deutschen Geistes haben darin ihren Hauptgrund, daß dort, in Frankreich, die thueresten Erinnerungen, Freiheiten, Gesetze und Überlieferungen der einzelnen Länder und Gebiete dem Ganzen weichen müssen; daß eine einzige Stadt monarchisch alle übrigen beherrscht, während in Deutschland, selbst bei den steinernen Communen, die heilige Unabhängigkeit für das alte Herzömmen, für das von Geschlecht zu Geschlecht überlieferte Gesetz, seiner Rücksicht und selbst seinem Vortheile weiden wollte *). Diese patriarchalische Treue für ihre provinziale Gestaltung, für ihre angerbten Localfretheite, 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 Württemberg 1816, wo die Opposition vom freisinnigen König, der eine ausgedehnte, die bisherigen Landesfreiheiten erhöhende Constitution geben wollte, sich entgegenstellte und auf die Einführung des Altebräuchlichen drang, und wo Uhland, der deutsche Sänger, das Lied dichtete:

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

Komponenten eines einfachen OCR-Workflows

1. Bildvorverarbeitung
- 2.
- 3.



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 aufzubauen? 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 erslang, 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 Vère, daß hier die Haare blond und dort schwarz sind! — — Aber die Rasse 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 weist.

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 Überlieferungen 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 heilige Anhänglichkeit für das alte Geschlehen, für das von Geschlecht zu Geschlecht überlieferte Gesetz, keiner Rücksicht und selbst keinem Vortheile weichen wollte *). Diese patriarchalische Treue für ihre provinzielle Gestaltung, 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 Württemberg 1816, wo die Opposition dem freisinnigen Könige, der eine ausgedehnte, die bisherigen Landesfreiheiten erhöhende Constitution geben wollte, sich entgegenstellte und auf die Einführung des Altagewöhnlichen drang, und wo Uhland, der deutsche Sänger, das Lied dichtete:

Wo je bei gutem, alten Wein
Der Württemberger zecht,
Da soll der erste Trunkspruch 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 aufzubauen? 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 erslang, 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 Vère, daß hier die Haare blond und dort schwarz sind! — — Aber die Rasse 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 weist.

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 Überlieferungen 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 heilige 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 Gestaltung, 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 Württemberg 1816, wo die Opposition dem freisinnigen Könige, der eine ausgedehnte, die bisherigen Landesfreiheiten erhöhende Constitution geben wollte, sich entgegenstellte und auf die Einführung des Altgebrauchlichen drang, und wo Uhland, der deutsche Sänger, das Lied dichtete:

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



Komponenten eines einfachen OCR-Workflows

1. Bildvorverarbeitung
2. Layoutanalyse
- 3.

15
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 auflebe? 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 erslang, 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 west.

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 Überlieferungen 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 heilige Unabhängigkeit 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 Gestaltung, für ihre angeerbten Vocalrechte, war zu allen Zeiten ein germanischer Charakterzug; er hat viel Unglück über Deutschland gebracht, aber

*¹) Um Ein Beispiel von Tausenden anzuführen, erinnern wir nur an die Versaifungskämpfe in Württemberg 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 deutsche Sänger, das Lied dichtete:

Wo je bei gutem, alten Wein
Der Württemberger zecht,
Der Wein ist erste Trunkspruch sein
Das gute alte Rept.

Komponenten eines einfachen OCR-Workflows

1. Bildvorverarbeitung
2. Layoutanalyse
3. Texterkennung

n5

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 auflebe? 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 erslang, 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 west.

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 Überlieferungen 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 heilige Unabhängigkeit für das alte Herkommen, für das von Geschlecht zu Geschlecht überlieferte Gesetz, keiner Rücksicht und selbst keinem Vortheile weichen wollte*).

*)) Um Ein Beispiel von Tausenden anzuführen, erinnern wir nur an die Versaifungskämpfe in Württemberg 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 deutsche Sänger, das Lied dichtete:

Wo je bei gutem, alten Wein
Der Württemberger zecht,
DRACM ist erste Trinkspruch sein
Das gute alte Rept.

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 auflebt? 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 bezicht, 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 wollte⁴⁾. Diese patriarchalische Treue für ihre provinzielle Gesitung, für ihre angeerbten Localrechte, war zu allen Zeiten ein germanischer Charakterzug; er hat viel Unglück über Deutschland gebracht, aber

⁴⁾ 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 Laudesfreiheiten erhöhende Constitution geben wollte, sich entgegenstellte und auf die Einstellung des Altgebrauchlichen drang, und wo Uhland, der deutsches Sänger das Lied dichtete:

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

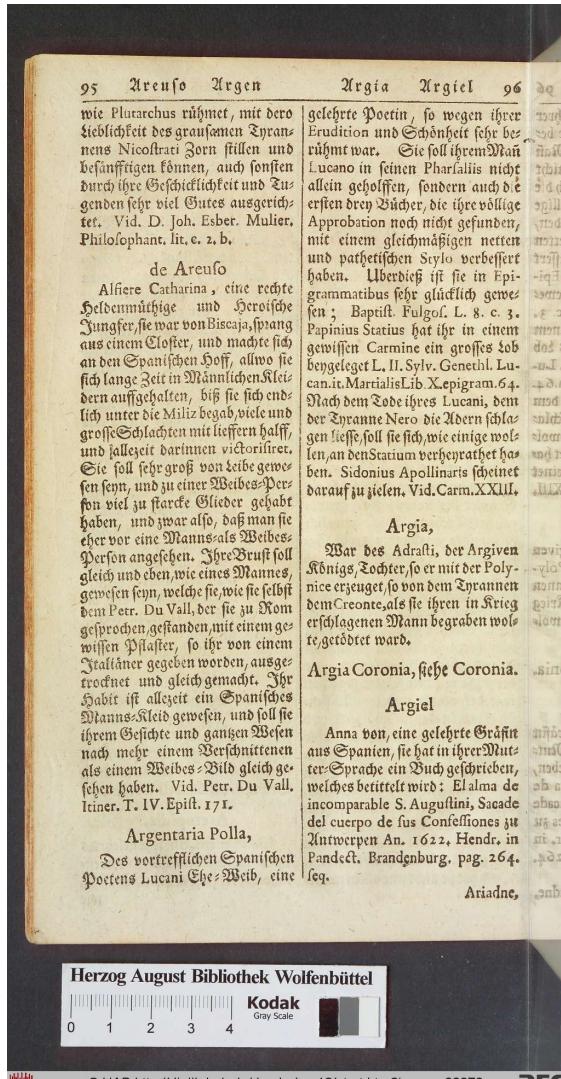


Komponenten eines OCR-Workflows: Bildvorverarbeitung

- 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 Argem Argia Argiel 96

wie Plutarchus rühmet, mit dero Lieblichkeit des grausamen Tyrannen Nicostrati Zorn stillen und besänftigen können, auch sonst durch ihre Geschicklichkeit und Tugenden sehr viel Gutes ausgerichtet. Vid. D. Joh. Esber. Mulier. Philosophant, lit. e. 2. b.

gelehrte Poetin, so wegen ihrer Eruditio[n] und Schönheit sehr berühmt war. Sie soll ihrem Man Lucano in seinen Pharsalis nicht allein geholfen, sondern auch die ersten drei Bücher, die ihre völlige Approbation noch nicht gefunden, mit einem gleichmäßigen netten und pathetischen Stylo verbessert haben. Überdies ist sie in Epigrammatibus sehr glücklich gewesen; Baptista Fulgos. L. 8. c. 3. Papinius Statius hat ihr in einem gewissen Carmine ein grosses Lob beigelegt L. II. Sylv. Genethl. Lucan. et. Marcialis Lib. Xepigram. 64. Nach dem Tode ihres Lucani, dem der Tyranne Nero die Adern schlagen ließ, soll sie sich, wie einige wollen, an den Statuum verheyrathet haben. Sidonius Apollinaris scheint darauf zu zielen. Vid. Carm. XXIII.

de Areuso
Altere Catharina, eine rechte Heldenmütige und Heroische Jungfer, sie war von Biscaya, sprang aus einem Closter, und mache sich an den Spanischen Hoff, alwo sie sich lange Zeit in Männlichen Kleidern aufzuhalten, bis sie sich endlich unter die Miliz begab, viele und grosse Schlachten mit liefern half, und jallezeit darinnen victorifret. Sie soll sehr groß von Leibe gewesen seyn, und zu einer Weibes-Person viel zu starke Glieder gehabt haben, und zwar also, daß man sie eher vor einem Mann als Weibes Person angesehen. Ihre Brust soll gleich und eben, wie eines Mannes, gewesen seyn, welche sie, wie sie selbst den Perr. Du Vall, der sie zu Rom gesprochen, gestanden, mit einem gewissen Pfaster, so ihr von einem Italiener gegeben worden, ausge trocknet und gleich gemacht. Ihr Habit ist allerzeit ein Spanisches Manns-Kleid gewesen, und soll sie ihrem Gesicht und ganzen Wesen nach mehr einem Verschnittenen als einem Weibes-Bild gleich gesehen haben. Vid. Perr. Du Vall, Itiner. T. IV. Epist. 171.

Argentaria Polla,
Des vortrefflichen Spanischen Poetens Lucani Ehe-Weib, eine

Argia,
War des Adrauli, der Argiven Königs Tochter, so er mit der Polynice erzeuget, so von dem Tyrannen dem Creonte, als sie ihren in Krieg erschlagenen Mann begraben wolle, getötet ward.

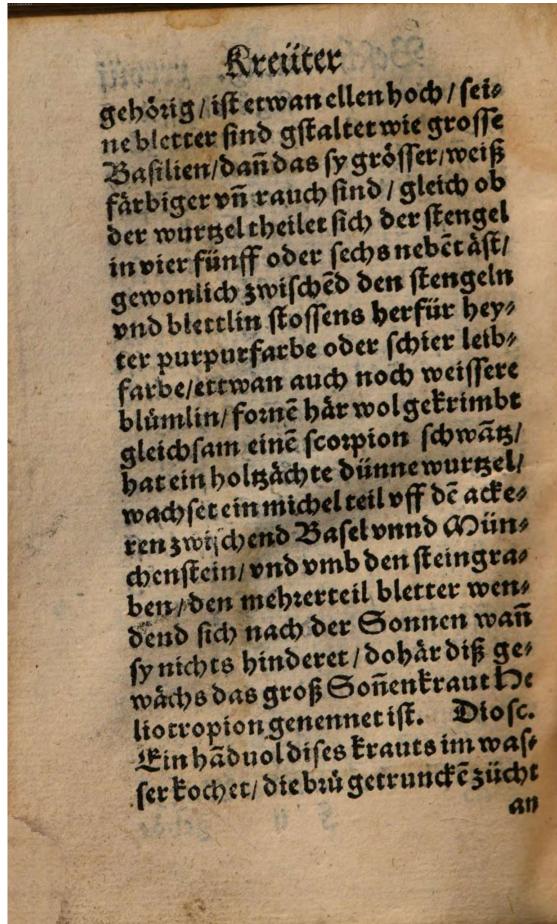
Argia Corona, siehe Corona.

Argiel
Anna von, eine gelehrte Gräfin aus Spanien, sie hat in ihrer Mutter-Sprache ein Buch geschrieben, welches betitelt wird: El alma de incomparable S. Augulini, Sacade del cuerpo de sus Confessiones zu Antwerpen An. 1622, Hendr. in Pandect. Brandenburg. pag. 264. seq.

Ariadne,



Komponenten eines einfachen OCR-Workflows: Deskewing

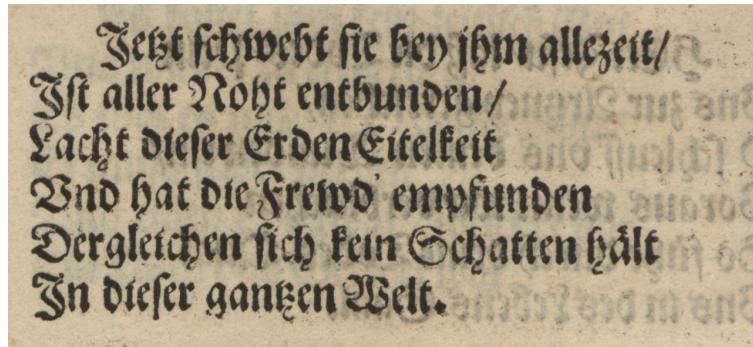


A digital representation of the same manuscript page, showing the text in a clear, modern font. The layout is identical to the original image, with the heading 'Kreüter' at the top and the detailed description of the plant below it.

Kreüter
gehörig / ist etwan ellen hoch / sei
ne bletter sind gſtalter wie grosse
Basilien / dañ das sy gröſſer / weiß
färbiger vñ rauch ſind / gleich ob
der wurgel theiler ſich der ſtengel
in vier fünff oder ſechs nebet äſt /
gewonlich zwischēd den ſtengeln
vnd bleetlin ſtoſſens herfür hey-
ter purpurfarbe oder ſchier leib-
farbe / etwan auch noch weiffere
blümlein / fornē hår wolgetrimbt
gleichsam eine scorpion schwāz /
hat ein holzachte dünnewurgel /
wachſet ein michel teil vff dē acke-
ren zwischend Basel vnd Mün-
ſchenſtein / vnd vmb den ſteingra-
ben / den mehrerteil bletter wen-
dend ſich nach der Sonnen wan-
n / sy nichts hinderet / dohár diſſ ge-
wächs das groſſ Sonen kraut He-
liotropion genennet iſt. Diosc.
Ein häduoldiſes Krauts im waſ-
ſer Kochet / die brū getrunkē zücht
an



Komponenten eines OCR-Workflows: Binarization



Jetzt schwebt sie bey ihm allezeit/
Ist aller Noht entbunden/
Lacht dieser Erden Eitelkeit
Und hat die Freyde empfunden
Dergleichen sich kein Schatten hält
In dieser ganzen Welt.



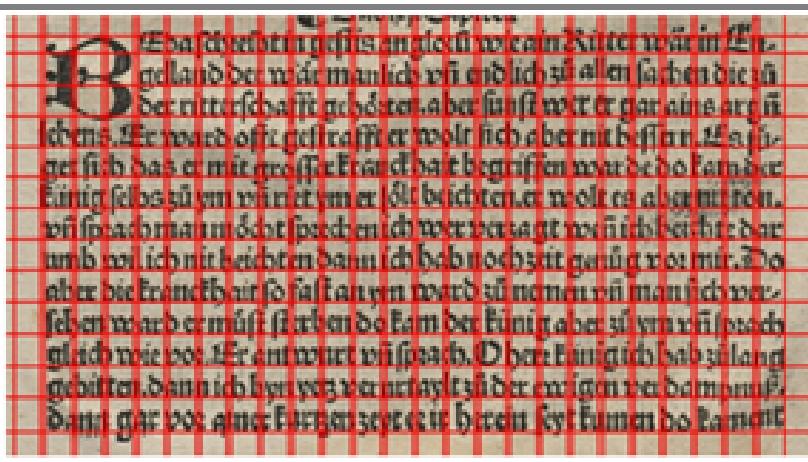
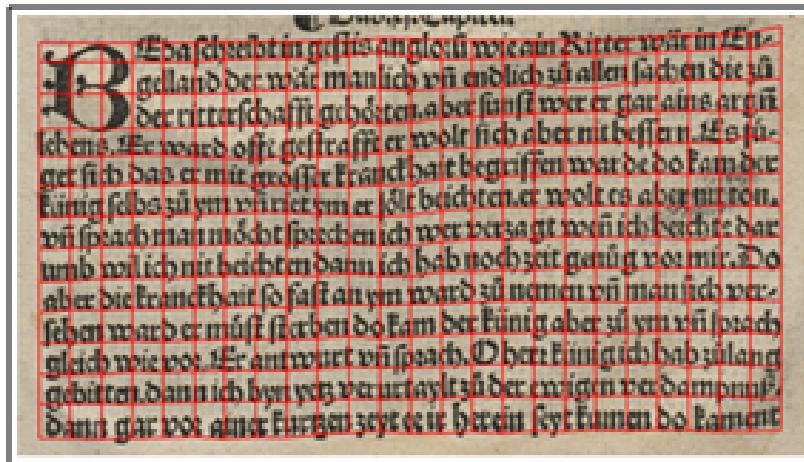
Komponenten eines OCR-Workflows: Despeckling

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 Eros². Dans cette série de huit couples divins, il en est un qui fait problème : Hermès-Hestia. Pourquoi les appairer ? 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-Eros), 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 Hermès³. 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 Eros². Dans cette série de huit couples divins, il en est un qui fait problème : Hermès-Hestia. Pourquoi les appairer ? 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-Eros), 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 alléger 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 Hermès³. De règle dans l'art plastique, l'association Hermès-Hestia



Komponenten eines einfachen OCR-Workflows: Dewarping



Komponenten eines OCR-Workflows: Einfluss der 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 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 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/
Die Fackel dieser Erden
Die Sonne/Kinder/Frennd' 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

SENS

UNIVERSITÄT · AKADEMIE · LIBRARY

Anfangsgründe der physischen Geographie. abb.

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

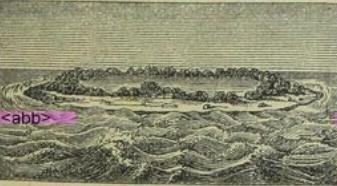


Fig. 5. Insel White-Sunday (Niedr. Inseln). abb.

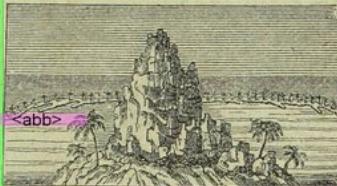


Fig. 6. Insel Bola Bola (Gesellschaftsinseln). abb.

da. 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. bb

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



Fig. 5. Insel White-Sundah (Niedr. Inseln). abb

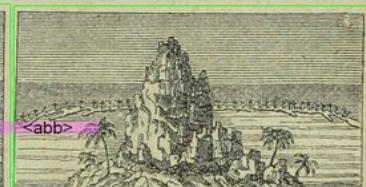


Fig. 6. Insel Bola Bola (Gesellschaftsinseln). abb

■ 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. bb

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

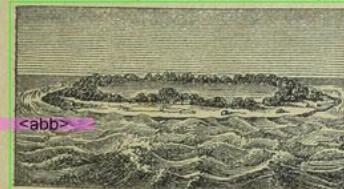


Fig. 5. Insel White-Sunday (Niedr. Inseln). abb

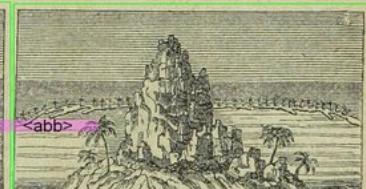


Fig. 6. Insel Bola Bola (Gesellschaftsinseln). abb

■ 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 et 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 zeilenorientiert 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
- Tesseract „Latin model“ (i.e. großmaßstäbliches Mehrsprachenmodell für Antiquaschriftarten): ca. 400 000 Zeilen in ca. 4 500 Schriftarten

Beyfuz̄

[Beyfuz̄](#)

Das Erſt Capitel

[Das Erſt Capitel](#)

Ariuosa Ampolata Brita

[Ariuosa Ampolata Brita](#)

nica Campanaria Metri

[nica Campanaria Metri](#)

caria mino: latīe. Melenoff Zan-

[caria minor latīe + Melenoff Zan-](#)

tes Thagetes Leptafelos Die

[tes Thagetes Leptafelos ¶ Die](#)

wirdigen maifter Auicenna Dia

[wirdigen maifter Auicenna Dia](#)

ſcorides beschreien vns vō diſem

[ſcorides beschreiben vns vō diſem](#)



Modelltraining: Trainingseffekte



Es kostet ihm kein zeitlich Gut
Bns wieder zu erwerben/
Es that es nicht der Opffer Blut/
Er musste selber sterben
Vnd einen Tod zwar / welcher gar
Ein Fluch vnd Grcwcl 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 ihm kein zeitlich Gut
Vns wieder zu erwerben/
Es that es nicht der Opffer Blut/
Er musste selber sterben
Vnd einen Tod zwar / welcher gar
Ein Fluch vnd Grcwcl 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 kostet ihm kein zeitlich Gut
Vns wieder zu terwerben/ Es
that es nicht der Opffer Blut/ Er
musste 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

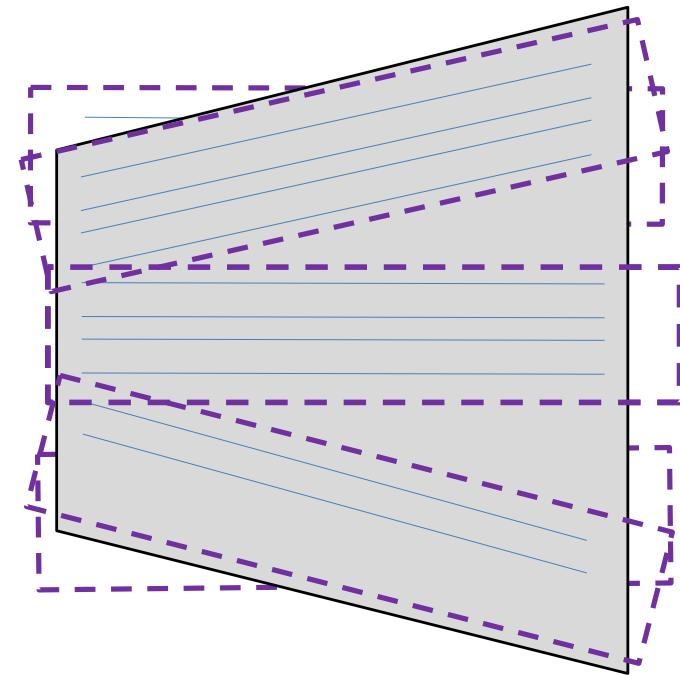
- historische Vorlagen bzw. ältere Digitalisate oftmals suboptimal für OCR
 - ▶ unterschiedliche **Beleuchtung**
 - ▶ charakteristische **Trapezform**
- verschiedene Bearbeitungsebenen
 - ▶ Dokument, Seite, Absatz (bzw. Textzone), Zeile
 - ▶ Operationen greifen **wiederholt** auf verschiedenen Ebenen 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)

Abbyy Finereader

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!

E u g e n i a.

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



OCR-Merging: Kotzebue „Schutzgeist“ (1814)

Abbyy Finereader

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!

E u g e n i a.

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

Tesseract

Eugeaia.

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



OCR-Merging: Kotzebue „Schutzgeist“ (1814)

Abbyy Finereader

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!

E u g e n i a.

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

Tesseract

Eugeaia.
O laß, H) mich. die Tht*ääea 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 haß, eh mich die Thränen ersticken,
Nur Einmal noch der Trennung Kuß
Auf die erblaßten Lippen drücken!
O gdne mir den letzten Genuß!



OCR-Merging: Kotzebue „Schutzgeist“ (1814)

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!

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*ääea 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 **haß**, eh mich die Thränen ersticken,
Nur Einmal noch der Trennung Kuß
Auf die erblaßten Lippen drücken!
O **gd**nne mir den letzten Genuß!



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

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

Abyy 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 ihm kein zeitlich Gut
Vns wieder zu erwerben/
Es that es nicht der Opffer Blut/
Er musste selber sterben
Vnd einen Tod zwar / welcher gar
EinFluch 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 musste selber sterben
Vnd einen Tod zwar / welcher gar
EinFliich'vud Grewel war.



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

Es kostet ihm kein zeitlich Gut
Vns wieder zu erwerben/
Es that es nicht der Opffer Blut/
Er muste selber sterben
Vnd einen Tod zwar / welcher gar
EinFluch 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 einen Tod zwar / welcher gar
EinFliich'vud 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.

OCRopus
Es kos**I**let jhm kein zeitlich Gut
Vns wieder zu **ter**werben/
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.

Tesseract
Es kostet jhm kein zeitlich Gut
Vns wieder zu erwerben/
Es ihaietz?i1ichi der Opffer Blui/
Er muſte selber ſterben
Vnd einen Tod zwar / welcher gar
Ein Fliich'vud 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.

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

**אָדִיר לְחַבּוֹרְתָנוּ אֵיכָה זוּה יְוָדָך בְּנֵי אָדָם,
אֲשֶׁר גָּדוֹתָם הַלְּבִיאָה בְּעֵבִי יִעַר עֲבַת, בְּחַרְשָׁה, 555**





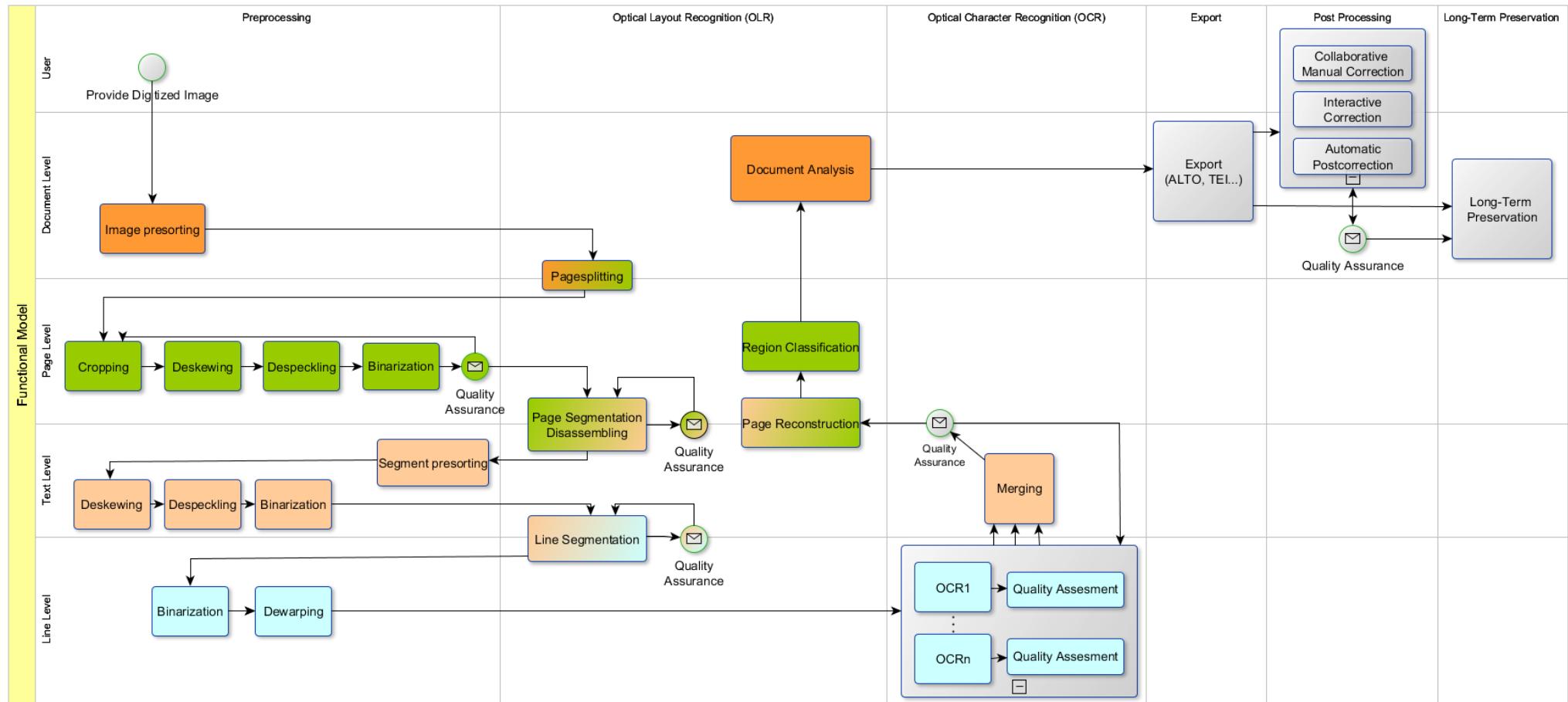
Komplexere OCR-Workflows

Komplexere OCR-Workflows

- „einfache“ OCR-Workflows in allen OCR-Lösungen implementiert
- **keine Möglichkeit** zur direkten Integration der diskutierten Optimierungsmöglichkeiten
- kein modulares **Workflowmanagementsystem** 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
 - ▶ Modeltraining
 - ▶ 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 ↪ 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

Tesseract Open Source OCR Engine (main repository)

theraysmith Initial push of one simple unittest

.github Make less verbose

android Update Android.mk

api Added AVX2 and AVX512 detector

arch Added ADAM optimizer, unless git screwed it up, cos there is no diff

ccmain Important fix to RTL languages saves last space on each line, which w...

ccstruct Added ADAM optimizer, unless git screwed it up, cos there is no diff

ccutil Define std::max under VS2017 x64

classify Fixed build broken by previous commits that added use of string in lo...

cmake Update SourceGroups.cmake

contrib helper script to generate dawg input files from text

cutil Remove extra semicolons



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>

