Vom Bild zum Text – praktische OCR für die DH

—

23.03.2021, 10–12 Uhr: Einführungsveranstaltung

Eventreihe

- **Dienstag, 23.03.2021, 10–12 Uhr: Einführungsveranstaltung**
- Mittwoch, 05.05.2021, 15–17 Uhr: OCR-D und OCR4all
- Mittwoch, 12.05.2021, 15–17 Uhr: Transkription, Training, Postcorrection
- Mittwoch, 19.05.2021, 15–17 Uhr: Hackathon
- Mittwoch, 15.09.2021, 14–16 Uhr: Abschlussveranstaltung

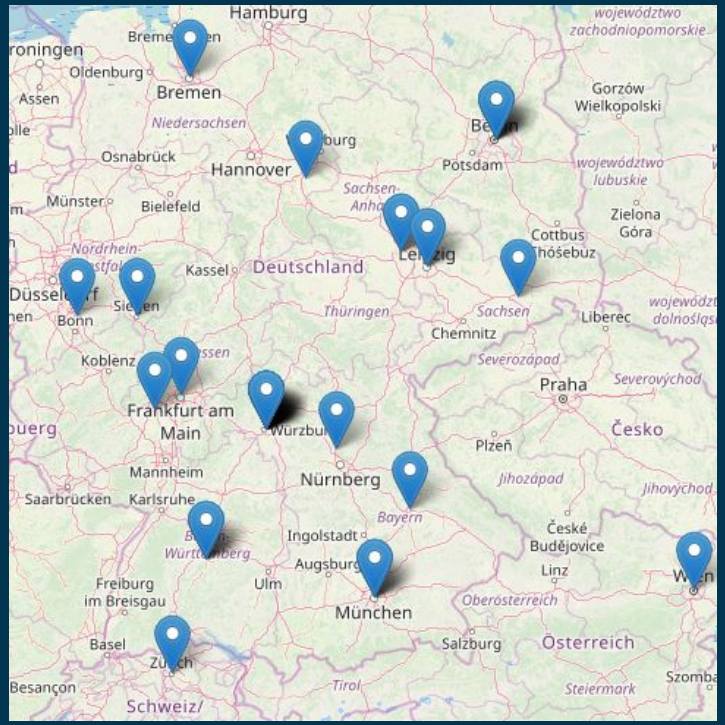
Agenda

- Vorstellung der DHd AG OCR
- Einführung in die OCR
- Projekte, Tools und Systeme
- Experimente

DHd AG OCR

Überblick

- AG im Sommer 2019 neu gegründet
- Derzeit 68 OCR-Interessierte in sehr heterogener Zusammensetzung:
  - Entwickler*innen
  - Vermittler*innen
  - Nutzer*innen
- Convenor:
  - Elisabeth Engl
  - Christian Reul



Ziele und Aktivitäten

- Vernetzung und Vermittlung zwischen spezifisch geistes-wissenschaftlichen Anforderungen und technischen Möglichkeiten
- Technologietransfer aus der Entwicklung in die Praxis
- Erarbeitung, Sicherung und Verbreitung von Best Practices
- Regelmäßige Treffen, Workshops und gemeinsame Projekte
  - Erstes DHd-gefördertes Projekt: Konvertierung vom “Entwicklerformat” PAGE-XML in “Nutzerformate” TEI, ALTO und PDF
  - Weihnachts- und Neujahrskolloquium “Einblicke und Ausblicke”
  - …

Kontakt

- DHd AG Info Seite: https://dig-hum.de/ag-ocr
- Mailingliste (Ankündigungen): https://lists.uni-wuerzburg.de/mailman/listinfo/ag-ocr
- Chat (alltäglicher Austausch): https://gitter.im/ag-ocr/community
- Homepage: https://dhd-ag-ocr.github.io
  - Materialien aus vergangenen Veranstaltungen
  - Institutionen und Projekte der Mitglieder
  - ...

Einführung in die OCR

Motivation

- Situation: Mehrheit aller jemals gedruckten Bücher gescannt und online frei zugänglich
- Problem: Scans erlauben nur ein eingeschränkte Nutzung
- ToDo: Herstellung hochwertiger maschinen-verarbeitbarer Texte:
  - sind durchsuchbar
  - sind annotierbar
  - können für zahlreiche quantitative Analyseverfahren verwendet werden
  - ...
- Methode: Optical Character Recognition (OCR)

Grundlagen und -begriffe – Workflow

Hauptkomponenten:

- Vorverarbeitung: Binarisierung, Geradestellen, …
- Segmentierung: Zerlegung in Regionen, Zeilen, (Zeichen); semantische Auszeichnung, Lesereihenfolge, …
- Texterkennung: Erkennung, Training, …
- Nachkorrektur: manuell, (semi-)automatisch

Grundlagen und -begriffe – Modelle und Training

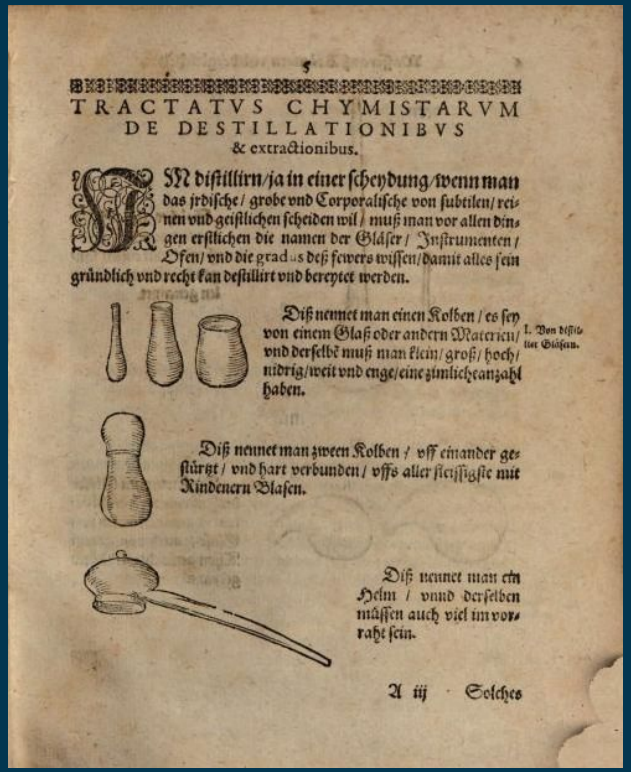
- Moderne OCR-Ansätze arbeiten hauptsächlich auf Zeilen- und nicht mehr auf Zeichenbasis
- Sog. “Modelle” extrahieren Text aus Textzeilenbildern
- Modelle müssen trainiert werden
  - Folgen Lernen-aus-Beispielen-Paradigma:
    1 Trainingsbeispiel = Zeilenbild + zugehörige Transkription
  - Nutzen Methoden des maschinellen Lernens
    (Neuronale Netze, Deep Learning, ...)
  - Generalisierbarkeit ↔ Genauigkeit
    - Gemischte/Polyfont/Omnifont Modelle
    - Werks-/Typenspezifische Modelle

Er wird eifrig geſammelt.

Fokus der Veranstaltung

- Open Source
- Historische Drucke
  - OCR für moderne Drucke bereits gelöst (?)
  - Historisches Material nach wie vor große Herausforderung
    - Schlechter Erhaltungs- und Druckzustand
    - Höchst komplexe Layouts
    - Große Varianz der verwendeten Drucktypen
    - Fehlende Standardisierung hinsichtlich Rechtschreibung
    - ...
  - Verwandtes Thema: Handwritten Text Recognition (HTR)
    - Zusätzliche Herausforderungen
    - Im Schnitt (deutlich) anspruchsvoller
    - Methodisch aber sehr ähnlich

Herausforderungen historischer OCR



TRACTATVS CHYMISTARVM
DE DESTILLATIONIBVS
& extractionibus.

M̃ distillirn/ja in einer scheydung/wenn man das jrdische/ grobe vnd Corporalische von subtilen/ reinen vnd geistlichen scheiden wil/ muß man vor allen dingen erstlichen die namen der Gläser/ Instrumenten/ Ofen/ vnd die gradus deß fewers wissen/ damit alles fein gründlich vnd recht kan destillirt vnd bereytet werden.

Diß nennet man einen Kolben/ es sey von einem Glaß oder andern Materien/ vnd derselbē muß man klein/ groß/ hoch/ nidrig/ weit vnd enge/ eine zimliche anzahl haben.

1. Von destillier Gläsern.

Diß nennet man zween Kolben/ vff einander gestürtzt/ vnd hart verbunden/ vffs aller sterckste mit Rindenern Blasen.

Diß nennet man ein Helm/ vnnd derselben müssen auch viel im vorrahte sein.

A iij Solches

u? v? n? un? nn? mi? tt?

brnst (brust)?

meüßerlin (modern: Mäusöhrlein)?

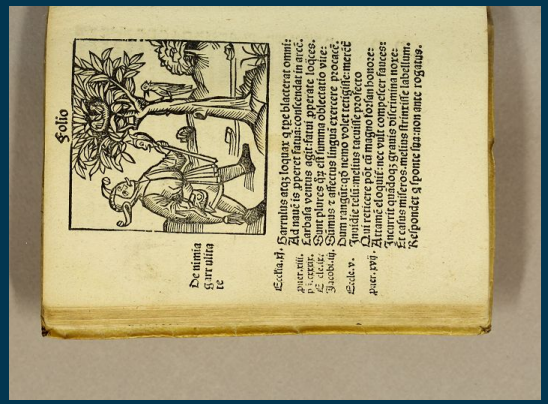
(Prof. Dr. Anke Lüdeling, HU Berlin)

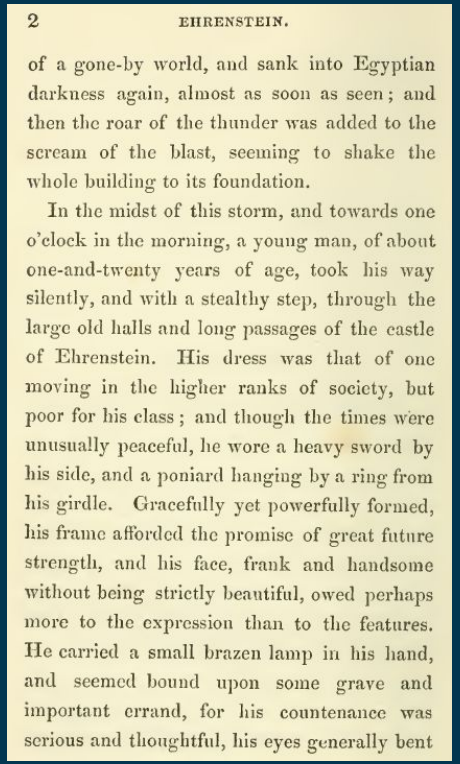
Einsatz von OCR sinnvoll? Und welche Lösung?

- Stark abhängig davon, wie das OCR Ergebnis genutzt werden soll!
- OCR wird sehr vielseitig eingesetzt → selten die eine, beste Lösung!
- Abhängig von vielen Faktoren:
  - Material: Womit habe ich es eigentlich zu tun?
  - Qualitätsanspruch der Nutzer*innen: Wie gut ist gut genug?
  - Aufwand: Wieviel Zeit/Geld kann/möchte ich investieren?
  - Weitere, äußere Constraints, z. B. verfügbare Hardware: Kann ich eine Lösung überhaupt sinnvoll anwenden?
  - ...

Material

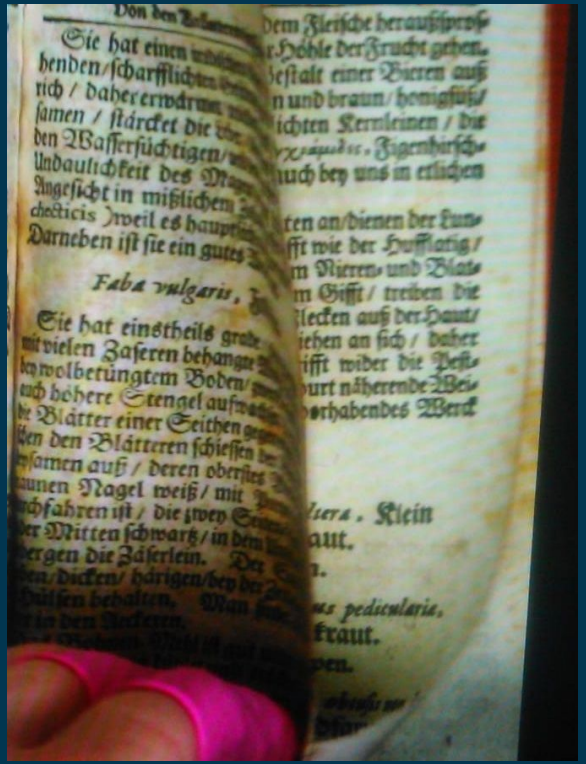
- Alter und Erhaltungszustand
- Druck- und Scanqualität
- Komplexität des Layouts
- Schrift und Sprache
- …




2 EHRENSTEIN.

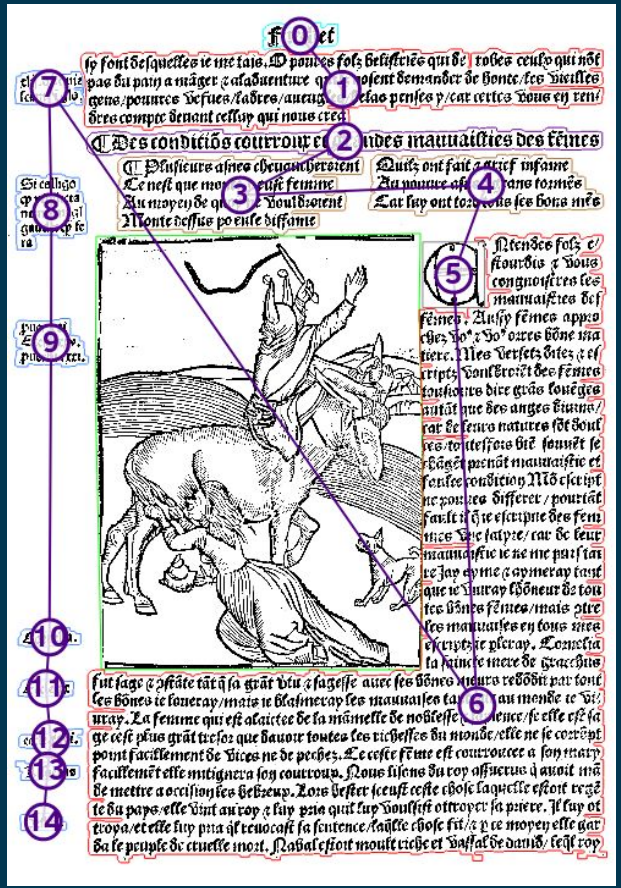
of a gone-by world, and sank into Egyptian darkness again, almost as soon as seen; and then the roar of the thunder was added to the scream of the blast, seeming to shake the whole building to its foundation.

In the midst of this storm, and towards one o'clock in the morning, a young man, of about one-and-twenty years of age, took his way silently, and with a stealthy step, through the large old halls and long passages of the castle of Ehrenstein. His dress was that of one moving in the higher ranks of society, but poor for his class; and though the times were unusually peaceful, he wore a heavy sword by his side, and a poniard hanging by a ring from his girdle. Gracefully yet powerfully formed, his frame afforded the promise of great future strength, and his face, frank and handsome without being strictly beautiful, owed perhaps more to the expression than to the features. He carried a small brazen lamp in his hand, and seemed bound upon some grave and important errand, for his countenance was serious and thoughtful, his eyes generally bent



Qualitätsanspruch

- Stark abhängig vom konkreten Anwendungsszenario!
- Text:
  - 100% Anspruch (z. B. für kritische Edition)
  - “wird schon passen” (Suche und quantitative Analyse?)
  - …
- Korrektheit und Detailtiefe der Segmentierung
  - Text/Nicht-Text-Trennung ausreichend?
  - Korrekte Lesereihenfolge nötig?
  - Semantische Auszeichnung? Was und wie?
  - …
- …



Zeit, Geld, Hardware, …

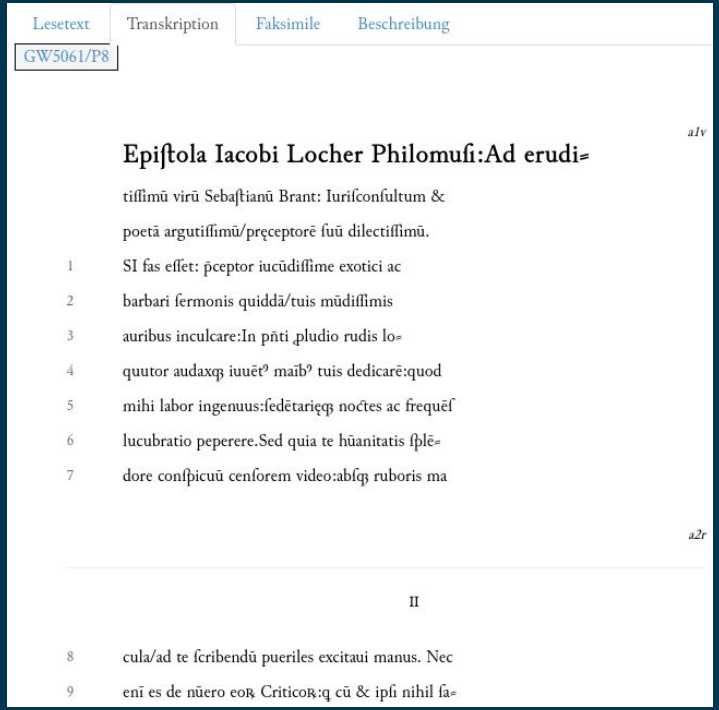
- Eigene Ressourcen
  - Zur Verfügung stehende finanzielle/personelle Mittel
  - Einarbeitungsaufwand
  - Deadlines
  - Hardware
  - …
- Tradeoff Qualität ↔ Laufzeit
  - Gerade in der Massenverarbeitung nicht zu verachten
  - Existierende Lösungen bieten diesbzgl. häufig bereits unterschiedliche Workflows, Modi, Modelle etc. an
- Ggf. Dienstleister-Support in Anspruch nehmen?

Konkrete Anwendungsbeispiele I

Kritische Edition am Beispiel *Narragonien digital* (http://narragonien-digital.de)

- Textqualität: 100% Anspruch auf sehr frühem Material (um 1500)
  - Manuelle Korrektur
  - Fortlaufendes werkspezifisches Training
- Segmentierung: Fehlerfreie Erfassung sämtlicher Layoutelemente

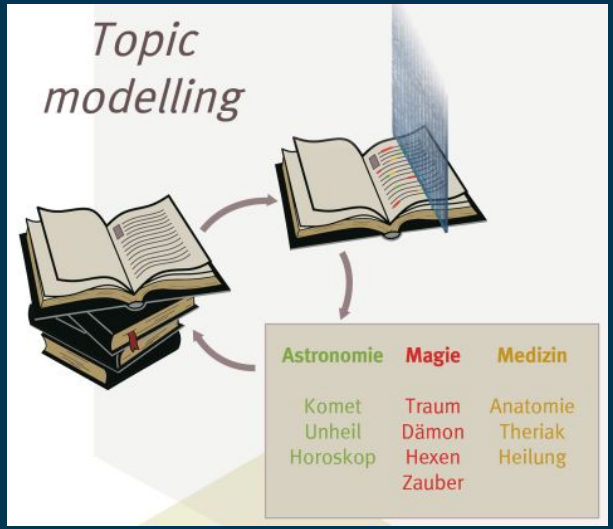
→ Hoher manueller Aufwand nötig …
… aber auch sinnvoll investiert



Konkrete Anwendungsbeispiele II

Quantitative Analyse am Beispiel Topic Modelling

- Task: Clustern von Wörtern/Begriffen zu Topics
- Wird normalerweise auf vergleichsweise umfangreichen Korpora ausgeführt
- Je nach Ansatz und Parametrisierung robust
  - ggü. textuellen OCR Fehlern
  - ggü. Segmentierungsfehlern, wie z. B. fehlende Spaltentrennung



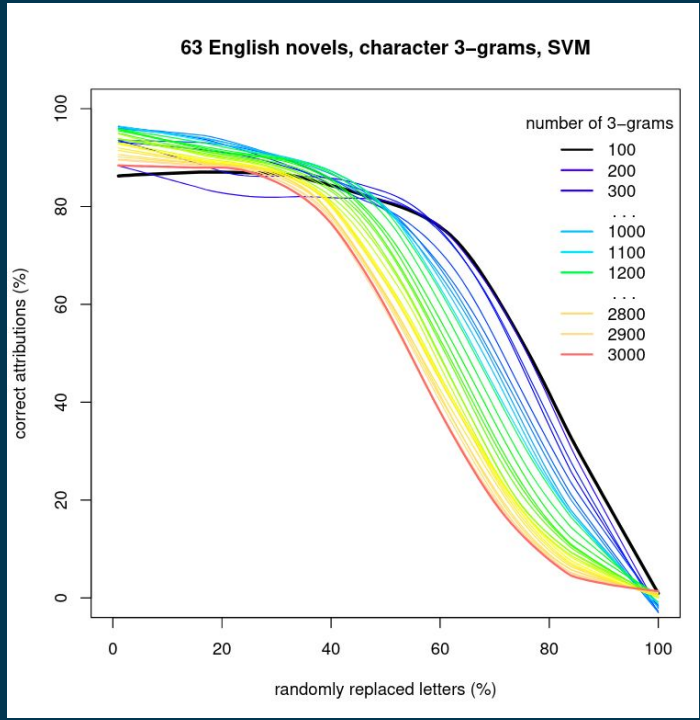
→ Fehlerhaftes OCR-Ergebnis keineswegs KO-Kriterium, allerdings viele Variablen

Konkrete Anwendungsbeispiele III

Quantitative Analyse am Beispiel Stilometrie

- Task: Zuordnung von Autor*innen, z. B. anhand der Worthäufigkeiten
- Vgl. Eder, M. (2013c). Mind your corpus: systematic errors in authorship attribution. Literary and Linguistic Computing, 28(4): 603–614.

→ Relativ robust gegen OCR Fehler

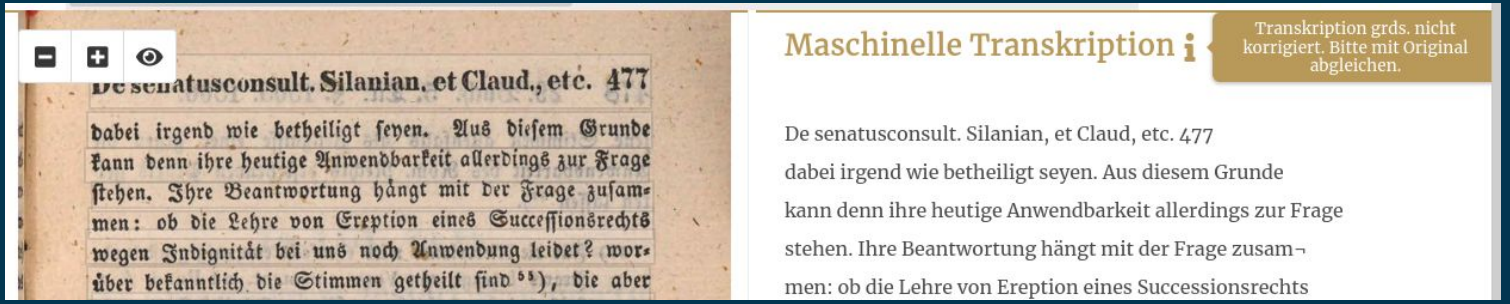


Konkrete Anwendungsbeispiele IV

(Unscharfe) Suche in massenhaft erschlossenen Korpora am Beispiel *iurisprudentia* (https://rwi.app/iurisprudentia/)

→ Tendenziell robust, tatsächlicher Nutzen aber abhängig von

- Qualität der textuellen OCR
- Korpusgröße, “Vollständigkeits-Anspruch”
- Robustheit der eingesetzten Suche
- Layoutfehler relevant? Suche nach Einzelbegriffen oder Wortketten?
- …



Projekte, Tools und Systeme

ABBYY

- etablierte, kommerzielle, proprietäre Desktopsoftware für OCR
- für Windows und Mac OS X
- als CLI-Version oder SDK für Linux
- nachtrainierbare Patterns (z. B. Sonderzeichen, Ligaturen…)

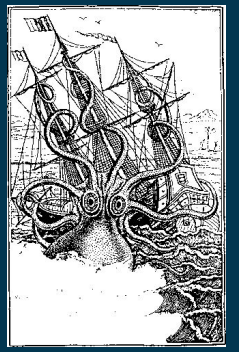
Transkribus

- Plattform für die AI-unterstützte Transkription von Dokumenten
- entstanden in EU-Projekten tranScriptorium und READ
- Kosten pro Seite zwischen 0,03€ und 0,19€
- als Webapp oder Client für diverse Betriebssysteme
- Modelle und Daten liegen beim Anbieter

Tesseract

- freie OCR-Lösung, *1985 bei HP, später Google
- in C++ implementiert
- mit Version 4 zeilenbasiert
- Modelle für über 100 Sprachen
- trainierbar mittels tesstrain

Ocropus, Kraken, Calamari



- LSTM-basierte OCR Engines, UI textbasiert
- in Python implementiert
- inzwischen auf GPUs nutzbar
- Ocropus *2007 DFKI/Google, aktuell Nvidia (später unter den Namen Ocropy, ocropy2, ocropus3, OCRopus4…)
- Kraken: „a turn-key OCR system optimized for historical and non-Latin script material“ – integriert in eScriptorium
- Calamari: Ocropus/Kraken Engine optimiert für größere Datenmengen und effizientes Training – integriert in OCR4all

OCR-D

- DFG-Projekt seit 2015 mit Ziel: Volltextdigitalisierung der VD
  - 2015–2017: Bestandsaufnahme
  - 2018–2020: Entwicklung von Prototypen, acht Satellitenprojekte
  - 2021–2023: Überführung in Produktivbetrieb, sieben Satellitenprojekte
- Einheitliche Konventionen, Interfaces und Implementierungen
- "Zerlegen" des OCR-Prozesses in seine Komponenten ("Prozessoren")
- Neben neuen Komponenten (typegroups_classifier, anybaseocr, ...) auch Wrapping existierender Tools (tesseract, calamari, ocr-fileformat...)
- Kommandozeilenbasiert (Fokus Massendigitalisierung)

OCR4all

- OCR-Workflow (nicht nur) für Frühdrucke als Komplettpaket mit Weboberfläche
- Segmentierung, GT-Produktion, Training, Erkennung, Nachkorrektur
- Für nicht-technische Anwender*innen konzipiert

Software für Digitale Bibliotheken

- Gängige Softwarelösungen für das Workflowmanagement und Präsentation in Bibliotheken und Archiven (Kitodo, Visual Library, Goobi, ...)
- Üblicherweise METS/MODS für Strukturdaten und ALTO für OCR
- Neuer Standard für die Präsentation: IIIF

⇒ Grundkenntnisse im Umgang mit METS helfen, um effektiv mit den Daten von Bibliotheken und Archiven zu arbeiten

Formate

- METS: LoC-Standard zum Beschreiben digitaler Objekte, insb. bibliografische Angaben und Strukturmetadaten wie Table of Content
- TEI: XML-Standard für die Digital Humanities. Extrem ausdrucksstark aber uneinheitlich und nicht ideal als OCR-Zielformat
- PAGE-XML: Gängiges Format für OCR und Segmentierungsergebnisse, de facto Standard für Competitions bei großen Konferenzen
- ALTO: LoC-Standard für OCR und Segmentierungsergebnisse, inzwischen weitgehend featuregleich zu PAGE-XML
- hOCR: Community-Standard basierend auf HTML, weniger featurereich, aber einfach zu handhaben
- TIFF/PNG: Verlustfreies Bildformat, zu bevorzugen wenn vorhanden
- JPEG2000: Ebenfalls verlustfrei, aber Support noch unzureichend

Experimente

Eventreihe

- Dienstag, 23.03.2021, 10–12 Uhr: Einführungsveranstaltung
- Mittwoch, 05.05.2021, 15–17 Uhr: OCR-D und OCR4all
- Mittwoch, 12.05.2021, 15–17 Uhr: Transkription, Training, Postcorrection
- Mittwoch, 19.05.2021, 15–17 Uhr: Hackathon
- Mittwoch, 15.09.2021, 14–16 Uhr: Abschlussveranstaltung

Ausgestaltung der weiteren Events

- Haben Sie bereits Vorkenntnisse im OCR-Bereich?
- Welche Themen sind Ihnen besonders wichtig?
- Haben Sie einen konkreten Use Case (Anwendungsfeld, Qualitätsansprüche, zu verarbeitende Materialien)?

Desiderate zur Volltextdigitalisierung der VD

- Wären Volltexte der VD-Titel für Sie hilfreich?
- Wie sollten diese in den VD bereitgestellt werden? (Qualität, Formate, ...)
- Für welche Arbeiten/Forschungsfragen würden Sie die VD-Volltexte gerne verwenden?

Vielen Dank für Ihre Aufmerksamkeit!