OCR und Strukturerkennung: Herausforderungen und Ansätze für die Zeitungsdigitalisierung

Clemens Neudecker (@cneudecker)

Staatsbibliothek zu Berlin – Preußischer Kulturbesitz

Virtueller 3. Workshop Retrodigitalisierung: „OCR – Prozesse und Entwicklungen“

1. März 2021

Staatsbibliothek
zu Berlin
Preußischer Kulturbesitz

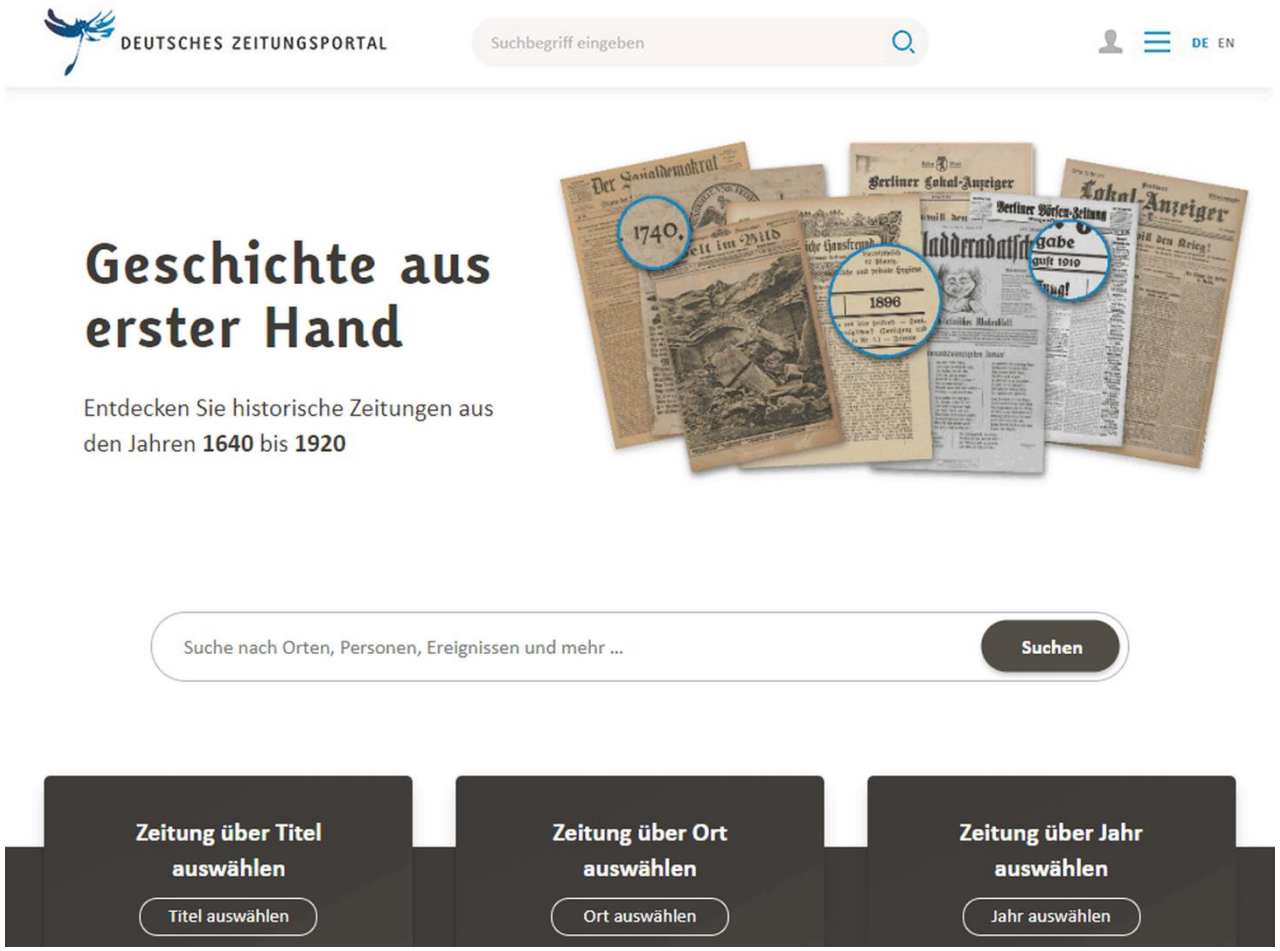
Who cares about yesterdays news?

- DDB Zeitungsportal
- DHd AG Zeitungen & Zeitschriften
- EuropeanaTech Insight Issue 16
- Dagstuhl Seminar 22361

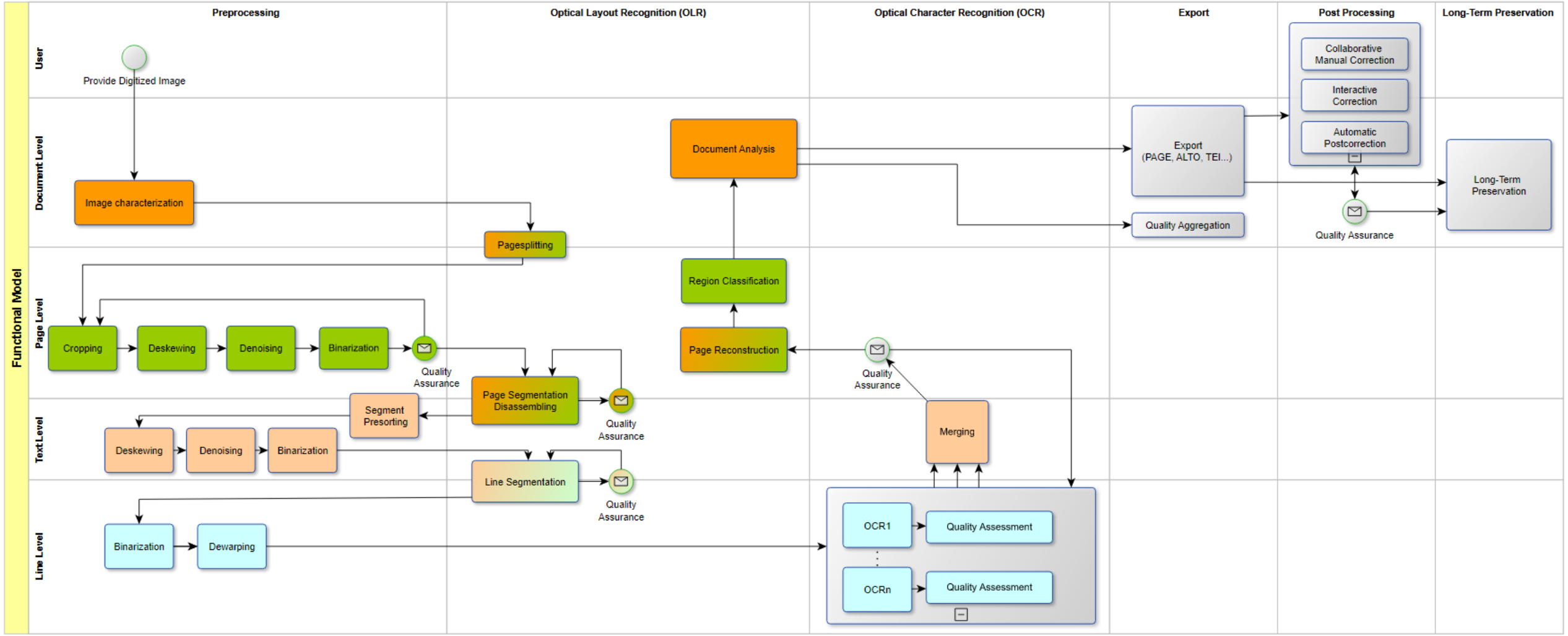
  „Computational Approaches for Digitized Historical Newspapers“
- Zahlreiche Digital Humanities Projekte, z.B.
  - *NewsEye* (EU)
  - *Oceanic Exchanges* (DiD)
  - *impresso* (SNF)

DDB Zeitungsportal

- Einheitliche Präsentation an einem Ort
- Wichtigste Funktionen für die Suche
  - Titelliste
  - Kalender
  - Volltextsuche
- „Fortgeschrittene Funktionen“ (Phase II)
  - Zitierbarkeit
  - Named Entities
  - Korpus Erstellung



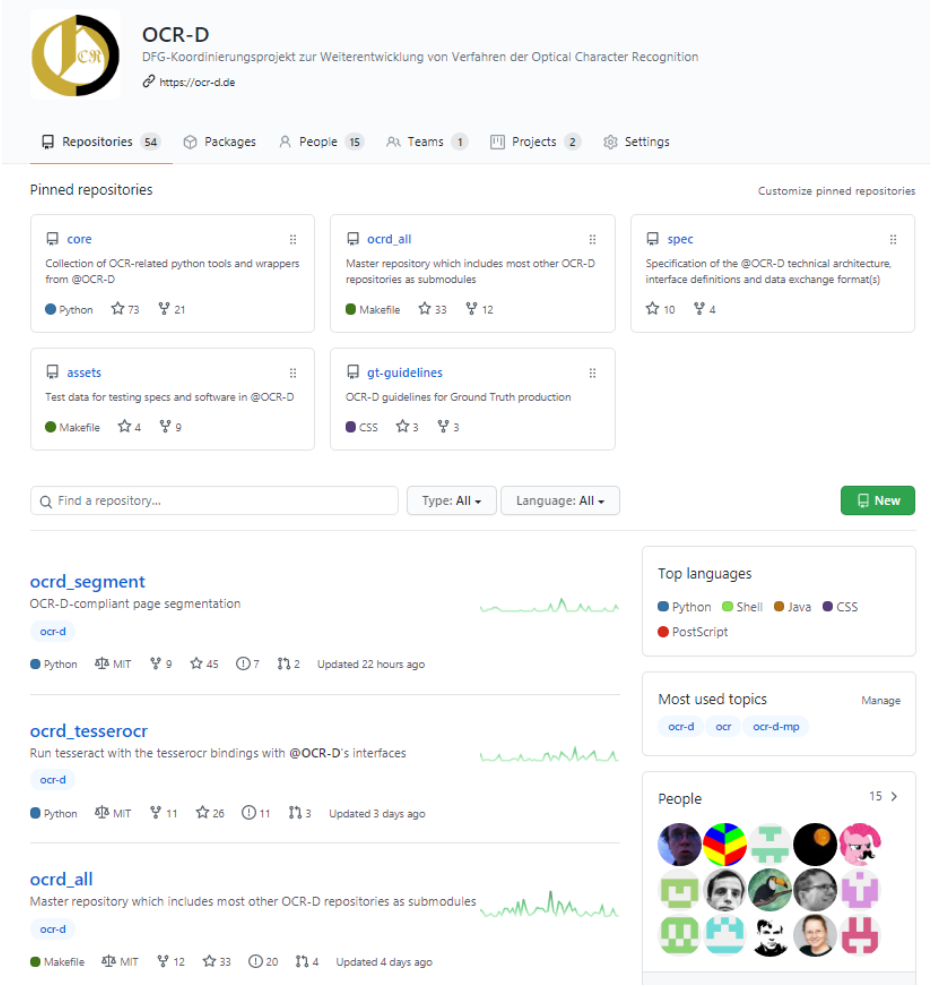
OCR Workflow



OCR-D

OCR-D
DFG-funded Initiative for Optical Character Recognition Development

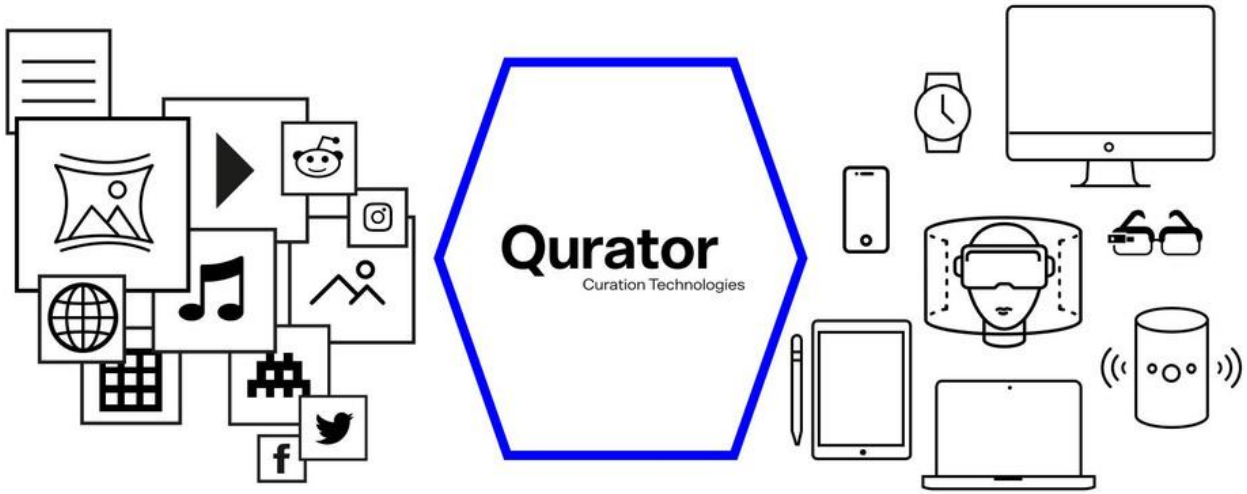
- Ziele: technische und organisatorische Grundlage für die OCR Verarbeitung der VD-Digitalisierungsprogramme schaffen
- Quelloffene und transparente Entwicklung
  - Spezifikationen & GT Richtlinien
  - Open source Software Tools
  - Wissens- und Erfahrungsaustausch in der Community
- 3 Phasen:
  - Phase I (2015-2018): Anforderungen erheben
  - Phase II (2018 – 2020): Entwicklung von Prototypen
  - Phase III (2021 – 2023): Implementierungen
- https://ocr-d.de



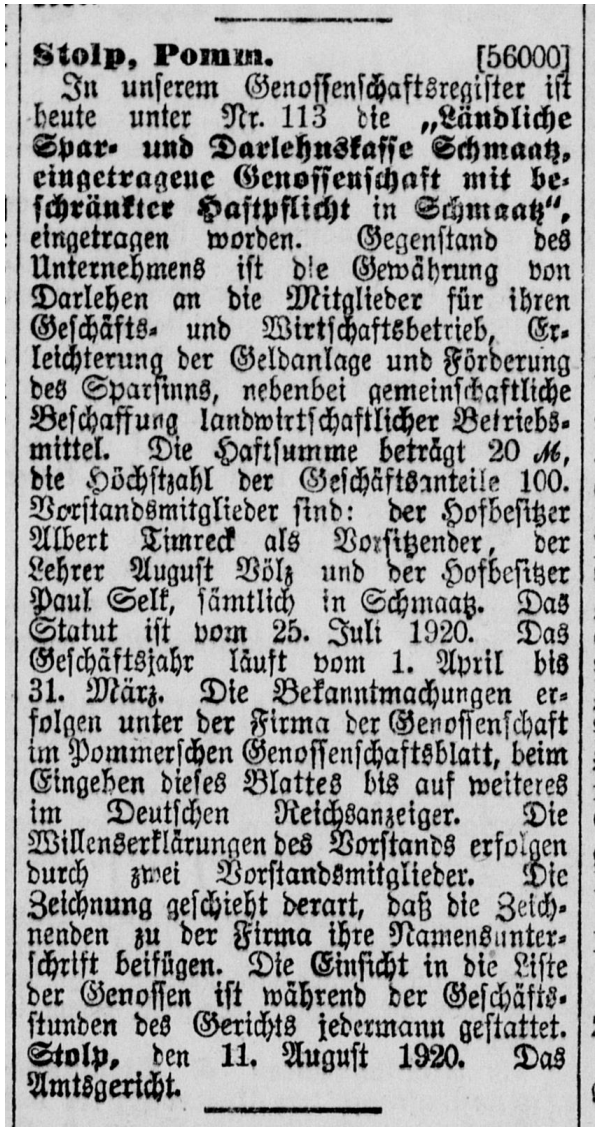
Qurator

Qurator
Curation Technologies

- Ziel: Technologien und Verfahren der Künstlichen Intelligenz für die Datenkuratierung nutzbar machen
- Use case: Digitalisiertes Kulturelles Erbe
- Entwicklung einer kompletten Pipeline:
  - Bildoptimierung
  - Binarisierung
  - Layout Analyse
  - OCR
  - OCR Nachkorrektur
  - Named Entity Recognition und Linking
  - Bildähnlichkeitssuche
- https://qurator.ai



OCR



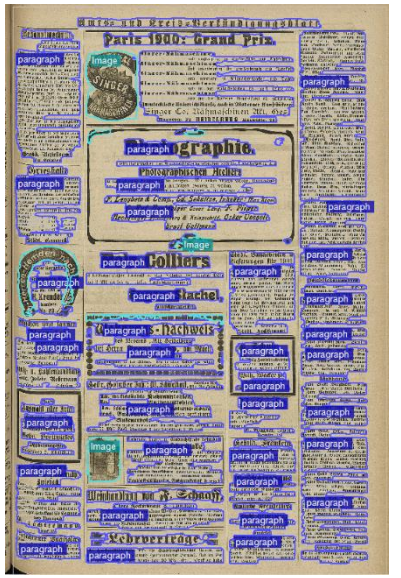
Stolp, Pomm. [56000]
In unſerem Genoſſenſchaftsregiſter iſt heute unter Nr. 113 die **„Ländliche Spar- und Darlehnskaſſe Schmaatz, eingetragene Genoſſenſchaft mit beſchränkter Haftpflicht** in Schmaatz“, eingetragen worden. Gegenſtand des Unternehmens iſt die Gewährung von Darlehen an die Mitglieder für ihren Geſchäfts- und Wirtſchaftsbetrieb, Erleichterung der Geldanlage und Förderung des Sparſinns, nebenbei gemeinſchaftliche Beſchaffung landwirtſchaftlicher Betriebsmittel. Die Haftſumme beträgt 20 M, die Höchſtzahl der Geſchäftsanteile 100. Vorſtandsmitglieder ſind: der Hofbeſitzer Albert Timreck als Vorſitzender, der Lehrer Auguſt Völz und der Hofbeſitzer Paul Selk, ſämtlich in Schmaatz. Das Statut iſt vom 25. Juli 1920. Das Geſchäftsjahr läuft vom 1. April bis 31. März. Die Bekanntmachungen erfolgen unter der Firma der Genoſſenſchaft im Pommerſchen Genoſſenſchaftsblatt, beim Eingehen dieſes Blattes bis auf weiteres im Deutſchen Reichsanzeiger. Die Willenserklärungen des Vorſtands erfolgen durch zwei Vorſtandsmitglieder. Die Zeichnung geſchieht derart, daß die Zeichnenden zu der Firma ihre Namensunterſchrift beifügen. Die Einſicht in die Liſte der Genoſſen iſt während der Geſchäftsſtunden des Gerichts jedermann geſtattet.
Stolp, den 11. Auguſt 1920. Das **Amtsgericht.**

Stolp, Pomm. [56000]
Jn unſerem Genoſſenſchaftsregiſter iſt heute unter Nr. 113 die ,,Ländliche Spar- und Darlehnskaſſe Schmaatz, eingetragene Genoſſenſchaft mit beſchränkter Haftpflicht in Schmaatz“, eingetragen worden. Gegenſtand des Unternehmens iſt die Gewährung von Darlehen an die Mitglieder für ihren Geſchäfts- und Wirtſchaftsbetrieb, Erleichterung der Geldanlage und Förderung des Sparſinns, nebenbei gemeinſchaftliche Beſchaffung landwirtſchaftlicher Betriebsmittel. Die Haftſumme beträgt 20 M, die Höchſtzahl der Geſchäftsanteile 100. Vorſtandsmitglieder ſind: der Hofbeſitzer Albert Timreck als Vorſitzender, der Lehrer Auguſt Völz und der Hofbeſitzer Paul Selk, ſämtlich in Schmaatz. Das Statut iſt vom 25. Juli 1920. Das Geſchäftsjahr läuft vom 1. April bis 31. März. Die Bekanntmachungen erfolgen unter der Firma der Genoſſenſchaft im Pommerſchen Genoſſenſchaftsblatt, beim Eingehen dieſes Blattes bis auf weiteres im Deutſchen Reichsanzeiger. Die Willenserklärungen des Vorſtands erfolgen durch zwei Vorſtandsmitglieder. Die Zeichnung geſchieht derart, daß die Zeichnenden zu der Firma ihre Namensunterſchrift beifügen. Die Einficht in die Liſte der Genoſſen iſt während der Geſchäftsſtunden des Gerichts jedermann geſtattet. Stolp, den 11. Auguſt 1920. Das Amtsgericht.

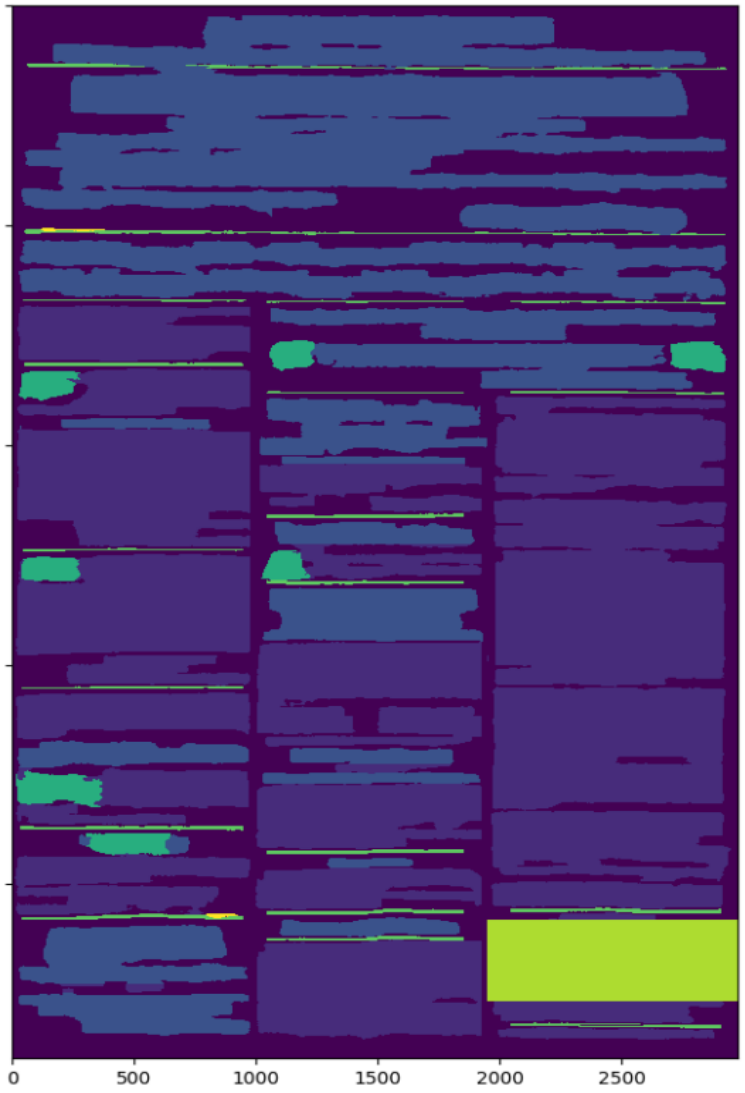
- Fehlerfreie OCR Resultate sind bspw. mit ocrd_calamari und einem auf dem Datensatz GT4HistOCR trainierten Modell möglich!
- Ein Vorteil von *Calamari* ggü. *Tesseract*, *OCRopus*, *Kraken*: Voting
- Deep Learning ermöglicht die Erkennung von sowohl Fraktur als auch Antiqua mit einem einzigen globalen und sprachunabhängigen Modell
- ABER…state-of-the-art OCR Engines benötigen für die Texterkennung bereits vorsegmentierte Textzeilen.

Layout Analyse

- Trainieren eines Machine Learning Systems (bspw. dhSegment) basierend auf einer *U-Net* / *ResNet* Architektur für pixel-wise Klassifikation anhand von Ground Truth Daten (mit Augmentation)
  - 1ste Iteration („reines ML“): gute Textzeilensegmentierung, aber Probleme mit Überschriften, Initialen, Reading Order
  - 2te Iteration („hybrid“): zusätzliche Heuristiken bringen substantielle Verbesserungen für sowohl Textzeilenerkennung als auch Reading Order
- Jedoch können selbst mit mehr GT Daten (noch) nicht alle Sonderfälle genügend abgedeckt werden

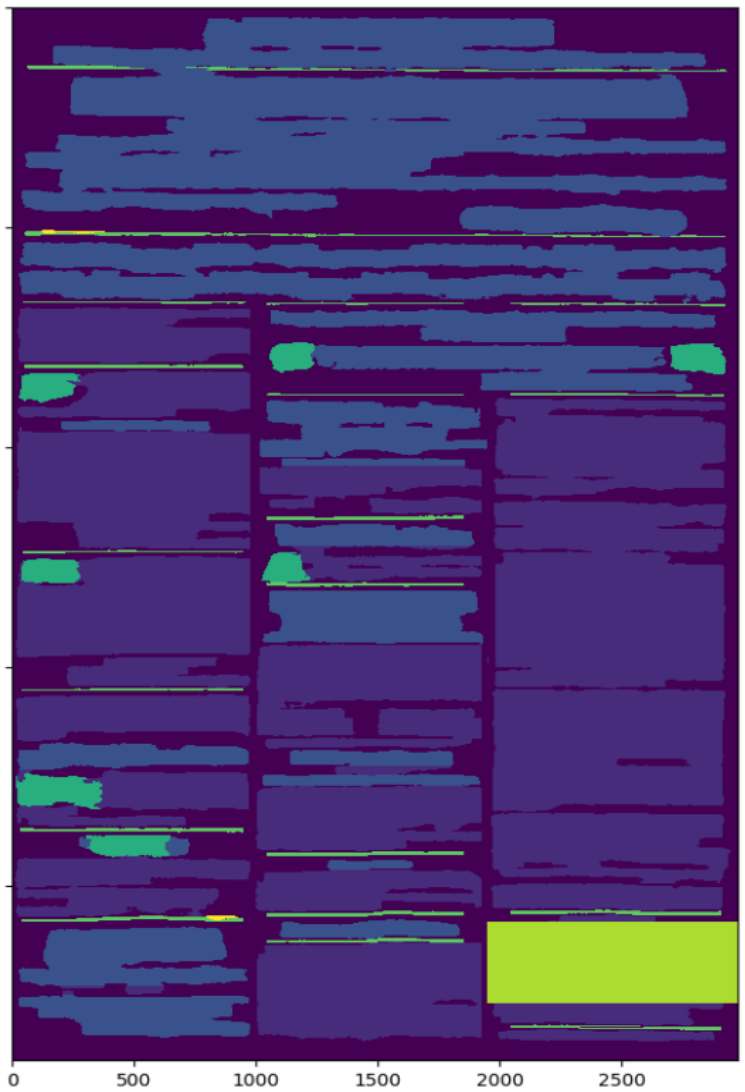


Layout Analyse

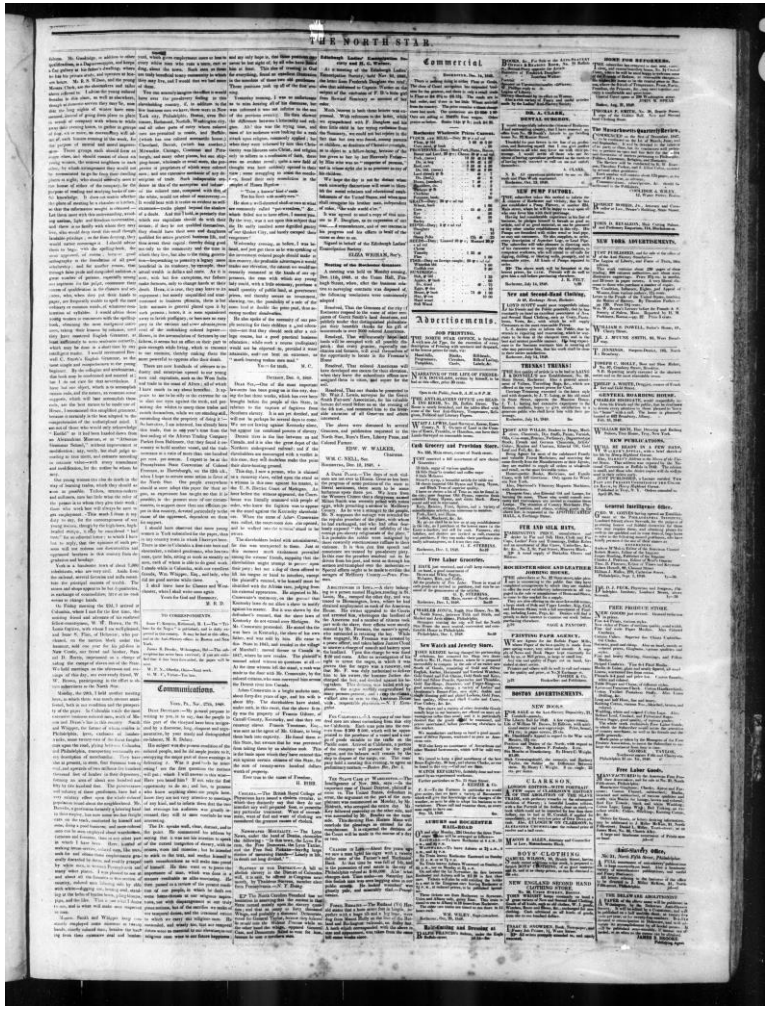


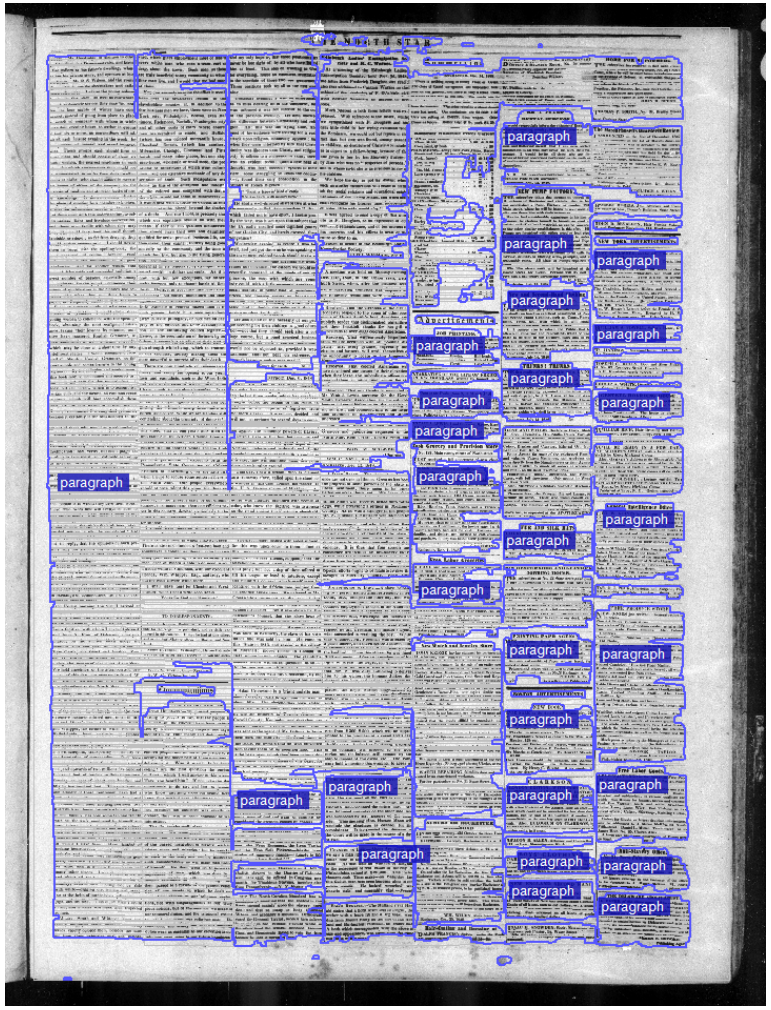
https://github.com/qurator-spk/eynollah

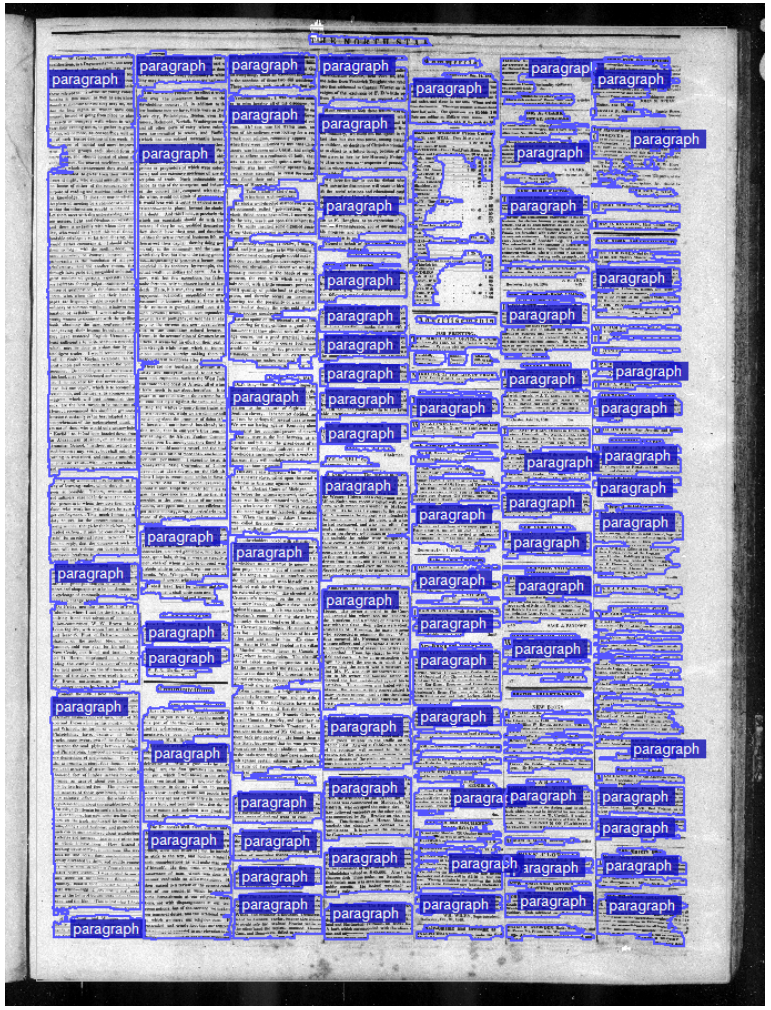
Reading Order



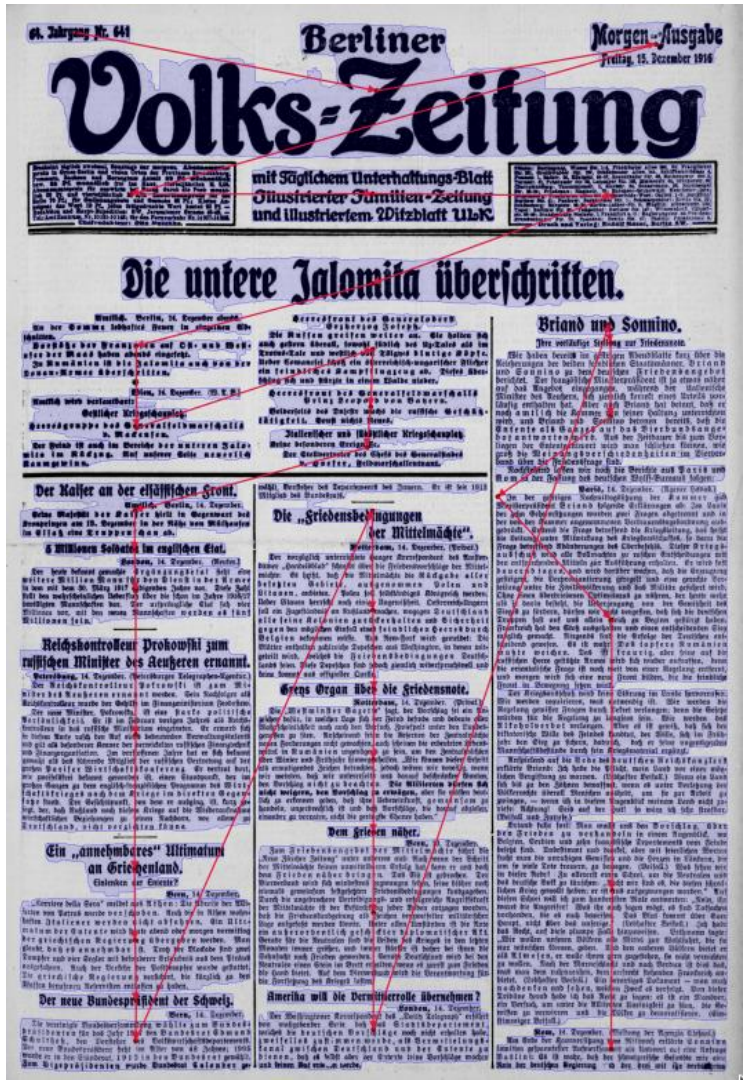
Herausforderungen

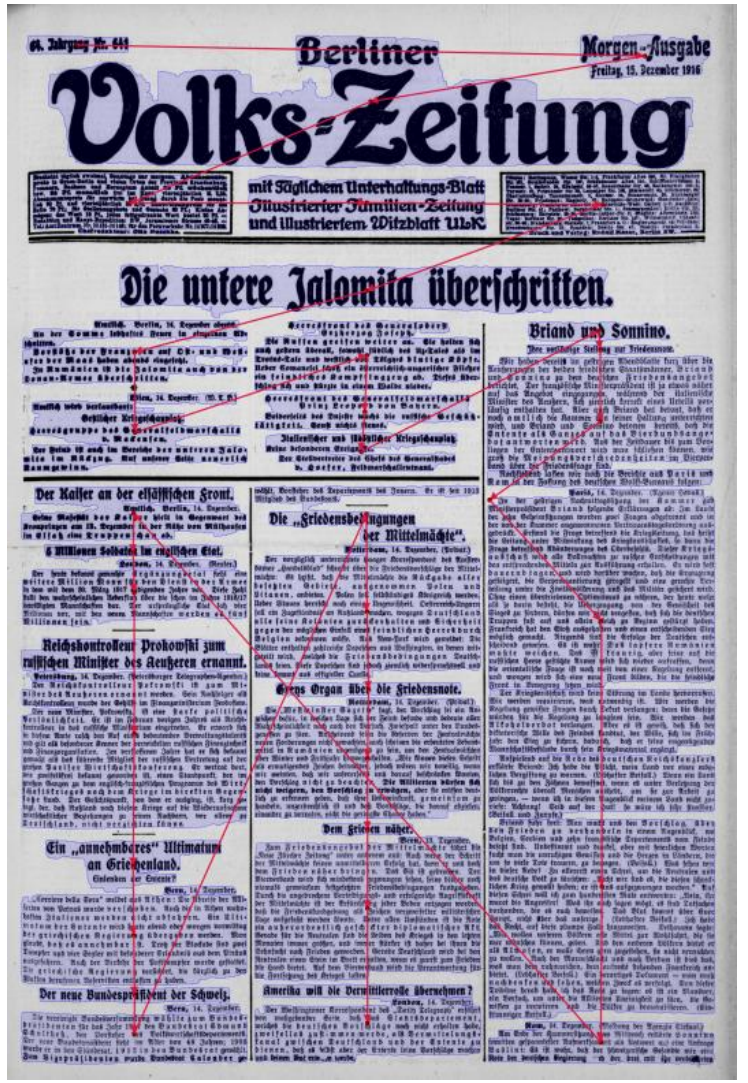




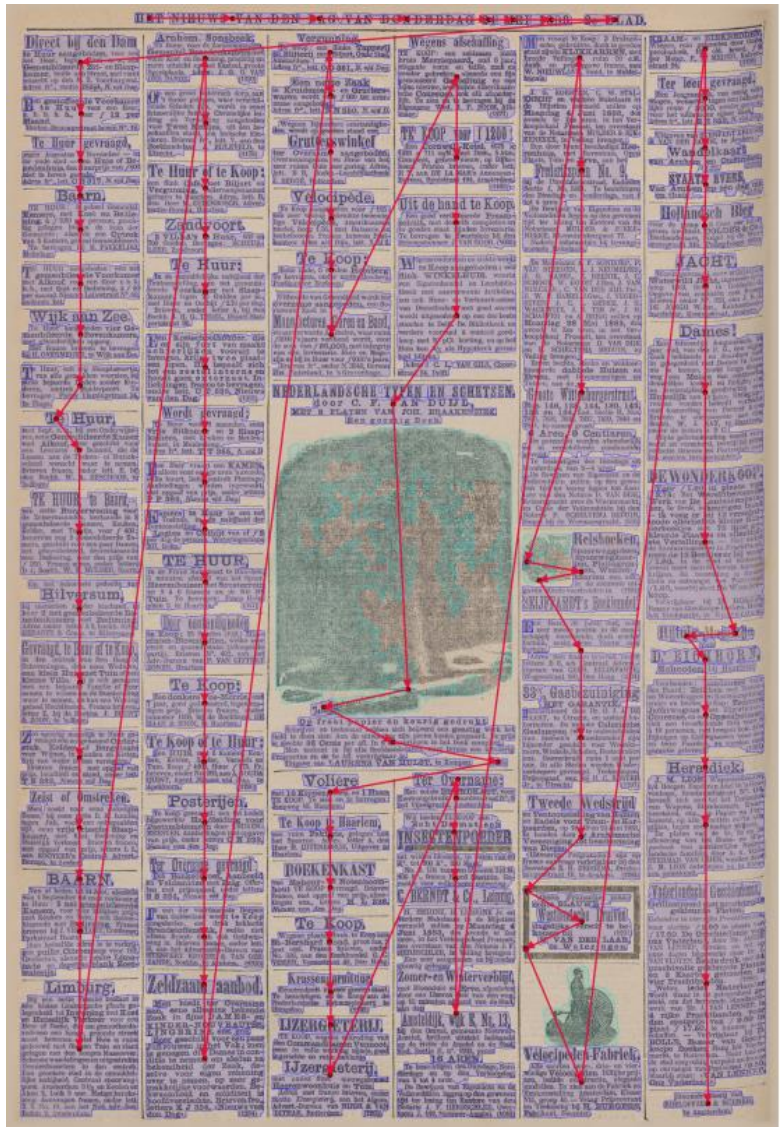


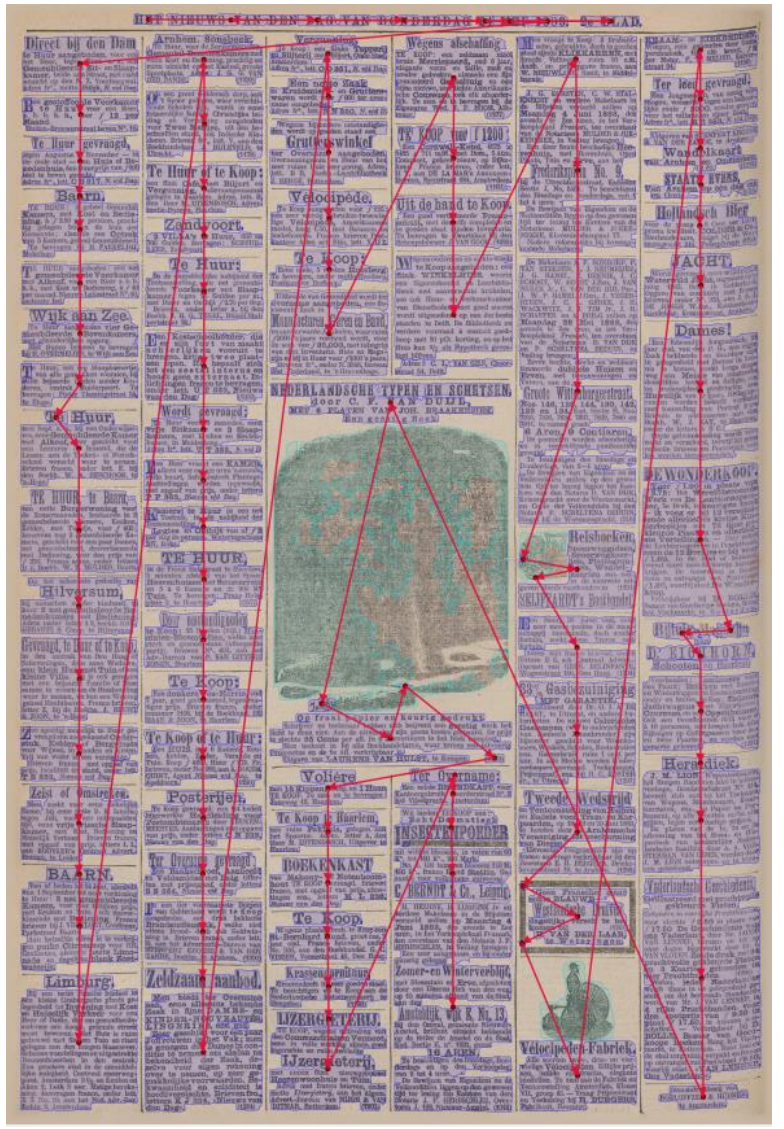
Herausforderungen





Herausforderungen

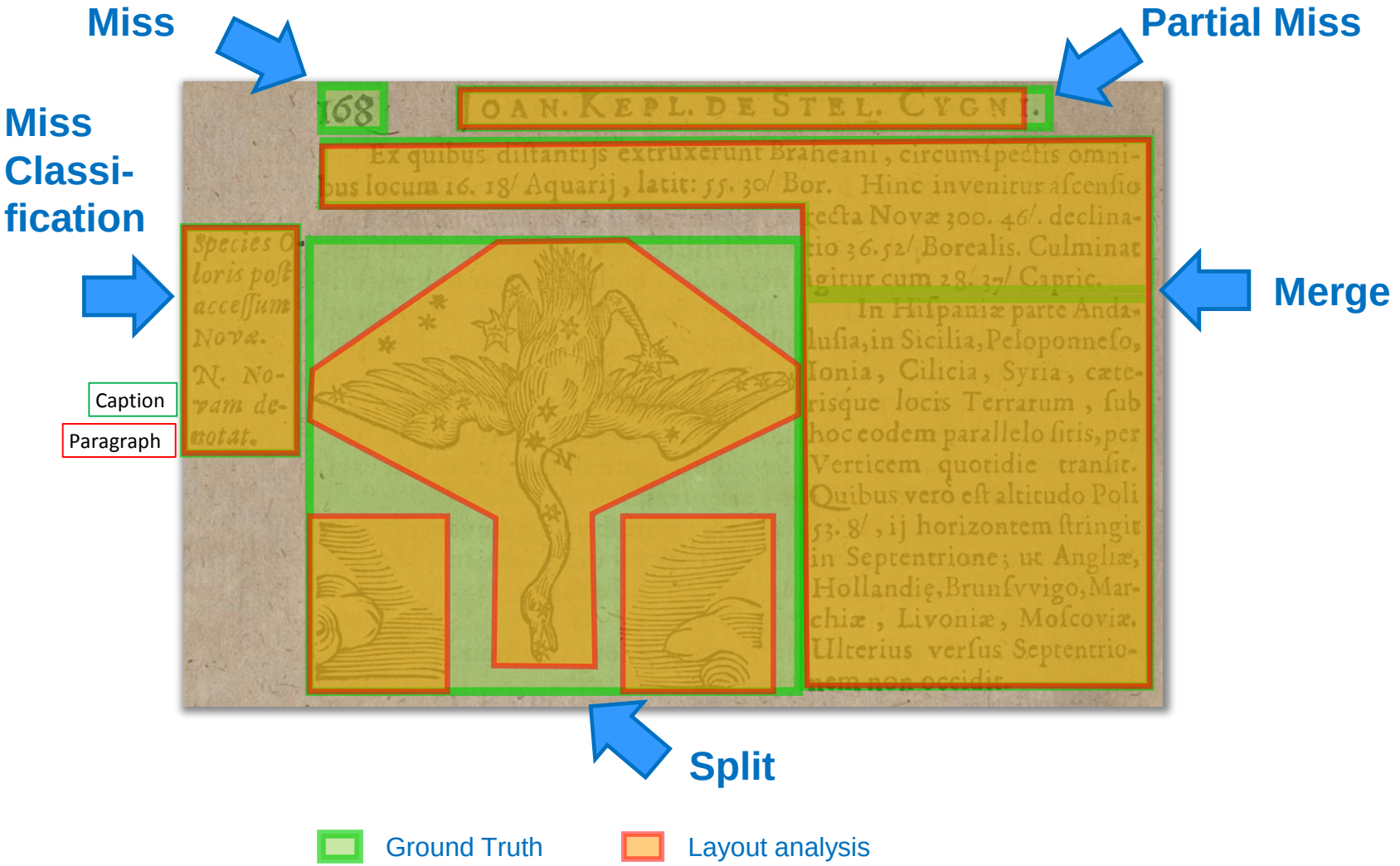




Evaluierung

Evaluierung

Evaluierung



$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

Ausblick – was fehlt uns noch zum Erfolg?

- Datensets historischer Zeitungen mit Layout GT
  a) von erheblichem Umfang (>1000 Seiten) und mit einer repräsentativen Abdeckung der Druckgeschichte
  b) mit granularen Annotationen für sämtliche relevanten Layout Elemente
  c) die offen zugänglich und frei nachnutzbar sind
- Methoden und Modelle für die Layoutanalyse die
  a) Computer Vision mit Natural Language Processing und
  b) Maschinelles Lernen mit Heuristiken in Balance bringen
- Community Standards und Empfehlungen für
  a) Metadaten für Layoutstrukturen und -elemente
  b) Metriken und Methoden für die Evaluierung von Layout Analyse

Danke für die Aufmerksamkeit!

Fragen?

Clemens Neudecker (@cneudecker)

Staatsbibliothek zu Berlin – Preußischer Kulturbesitz

Virtueller 3. Workshop Retrodigitalisierung: „OCR – Prozesse und Entwicklungen“

1. März 2021

Staatsbibliothek
zu Berlin
Preußischer Kulturbesitz