

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

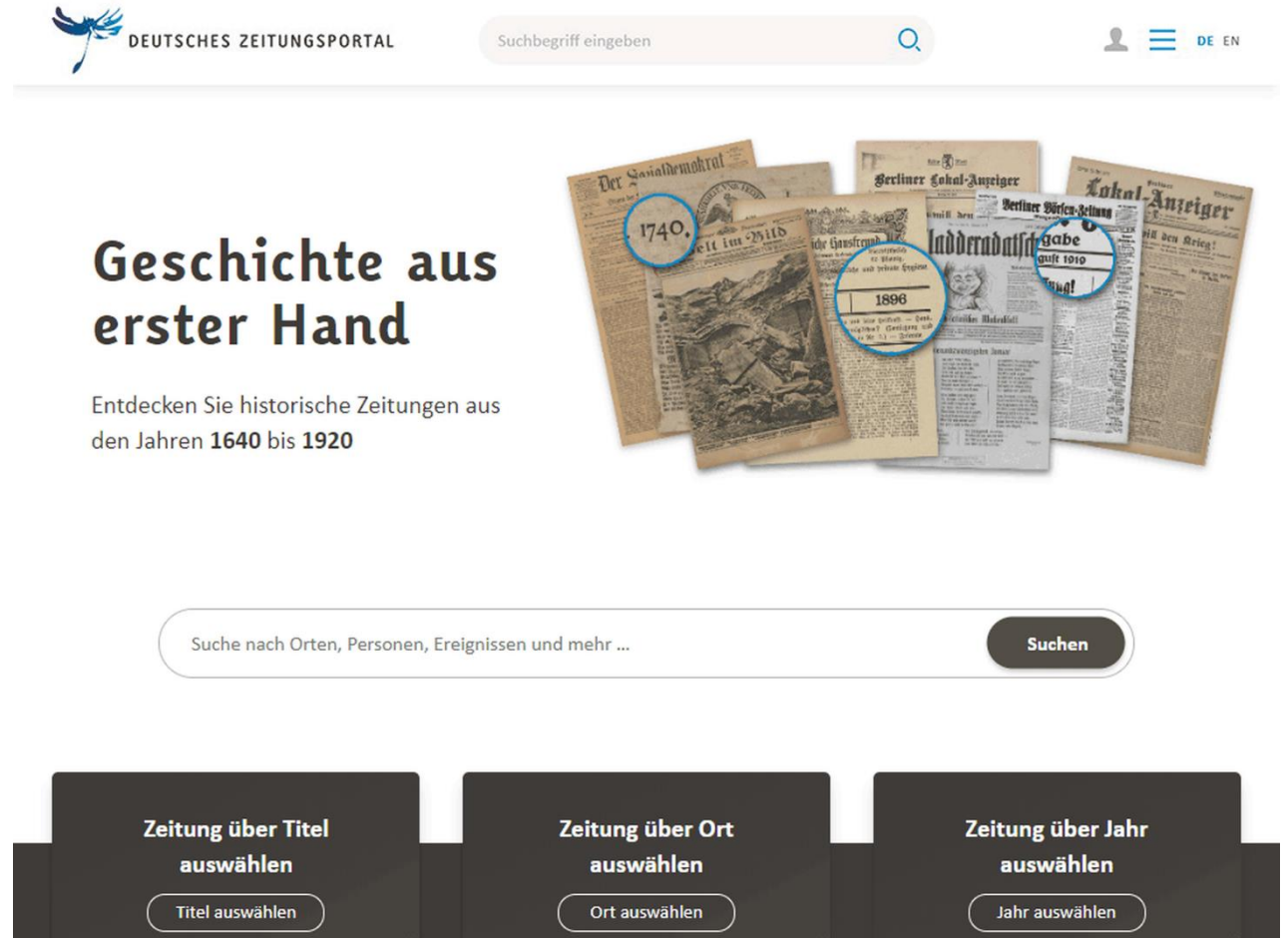
Who cares about yesterdays news?

- [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



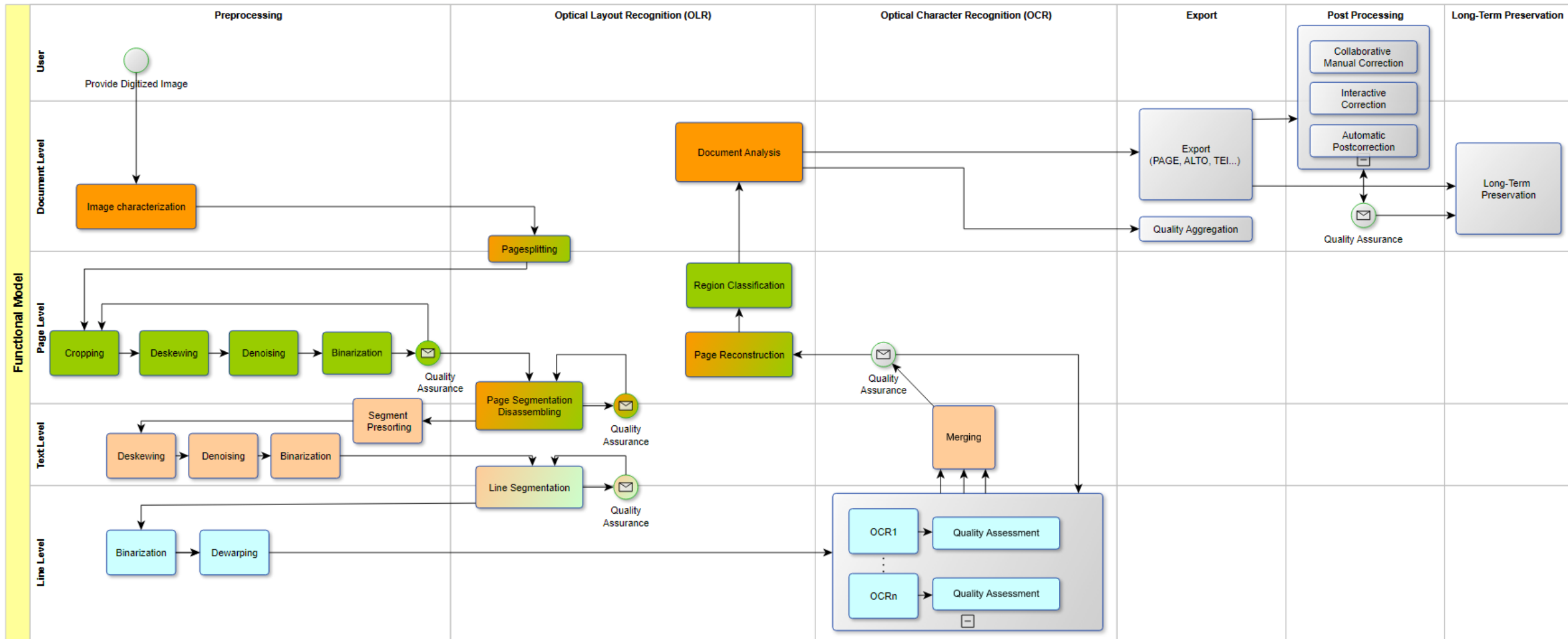
The screenshot shows the homepage of the DDB Zeitungsportal. At the top left is the logo of a blue dragon and the text "DEUTSCHES ZEITUNGSPORTAL". To the right is a search bar with the placeholder text "Suchbegriff eingeben" and a magnifying glass icon. Further right are icons for a user profile, a menu, and language options "DE EN".

The main content area features a large heading "Geschichte aus erster Hand" and a sub-heading "Entdecken Sie historische Zeitungen aus den Jahren 1640 bis 1920". To the right of this text is a collage of historical newspaper pages, including "Der Sozialdemokrat" (1740), "Berliner Lokal-Anzeiger", and "Lokal-Anzeiger" (1896). The collage also includes a page with the date "1896" and another with "1910".

Below the main content is a search bar with the placeholder text "Suche nach Orten, Personen, Ereignissen und mehr ..." and a "Suchen" button.

At the bottom, there are three dark grey buttons with white text: "Zeitung über Titel auswählen" (with a sub-button "Titel auswählen"), "Zeitung über Ort auswählen" (with a sub-button "Ort auswählen"), and "Zeitung über Jahr auswählen" (with a sub-button "Jahr auswählen").

OCR Workflow



OCR-D

- 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>



OCR-D
DFG-Koordinierungsprojekt zur Weiterentwicklung von Verfahren der Optical Character Recognition
<https://ocr-d.de>

Repositories 54 Packages 15 People 15 Teams 1 Projects 2 Settings

Pinned repositories

- core**: Collection of OCR-related python tools and wrappers from @OCR-D. Python, 73 stars, 21 forks.
- ocrd_all**: Master repository which includes most other OCR-D repositories as submodules. Makefile, 33 stars, 12 forks.
- spec**: Specification of the @OCR-D technical architecture, interface definitions and data exchange format(s). 10 stars, 4 forks.
- assets**: Test data for testing specs and software in @OCR-D. Makefile, 4 stars, 9 forks.
- gt-guidelines**: OCR-D guidelines for Ground Truth production. CSS, 3 stars, 3 forks.

Find a repository... Type: All Language: All New

ocrd_segment
OCR-D-compliant page segmentation
ocr-d
Python MIT 9 stars 45 forks 7 issues 2 pull requests Updated 22 hours ago

ocrd_tesseract
Run tesseract with the tesseract bindings with @OCR-D's interfaces
ocr-d
Python MIT 11 stars 26 forks 11 issues 3 pull requests Updated 3 days ago

ocrd_all
Master repository which includes most other OCR-D repositories as submodules
ocr-d
Makefile MIT 12 stars 33 forks 20 issues 4 pull requests Updated 4 days ago

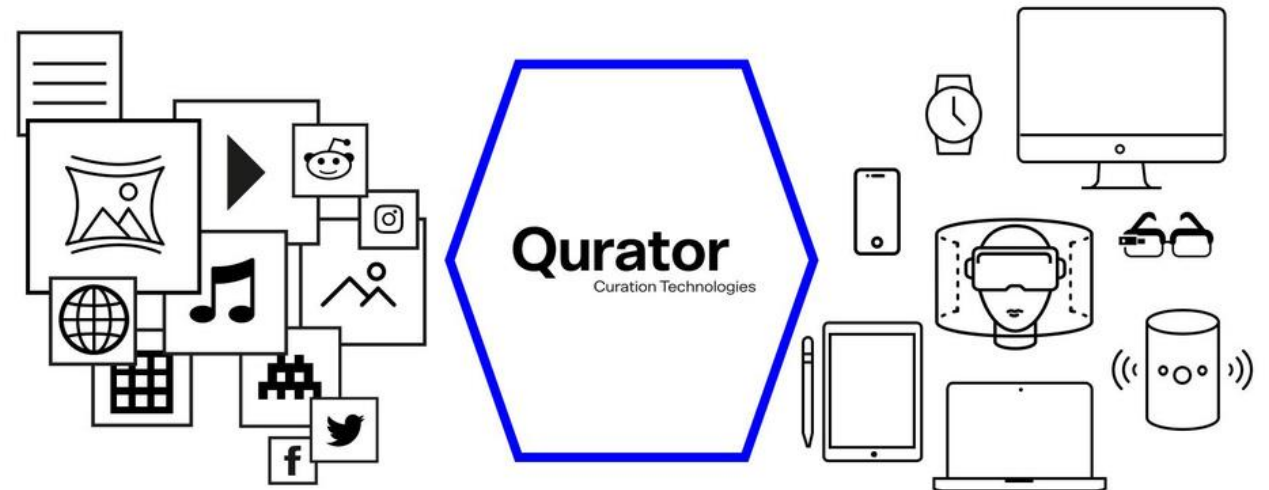
Top languages
Python Shell Java CSS PostScript

Most used topics
ocr-d ocr ocr-d-mp

People 15 >

Qurator

- 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

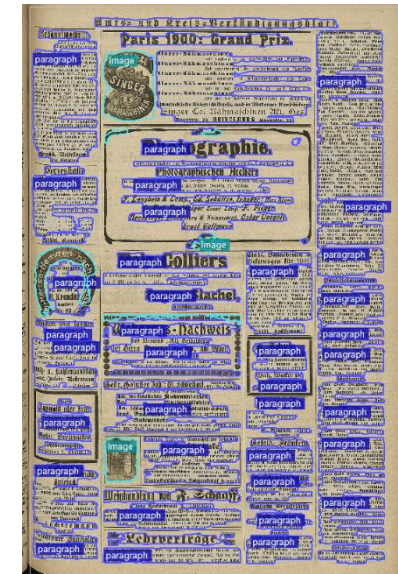
Stolp, Pomm. [56000]
In unserem Genossenschaftsregister ist heute unter Nr. 113 die „Ländliche Spar- und Darlehnskasse Schmaatz, eingetragene Genossenschaft mit beschränkter Haftpflicht in Schmaatz“, eingetragen worden. Gegenstand des Unternehmens ist die Gewährung von Darlehen an die Mitglieder für 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 Pommerischen 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]
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 Pommerischen 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.

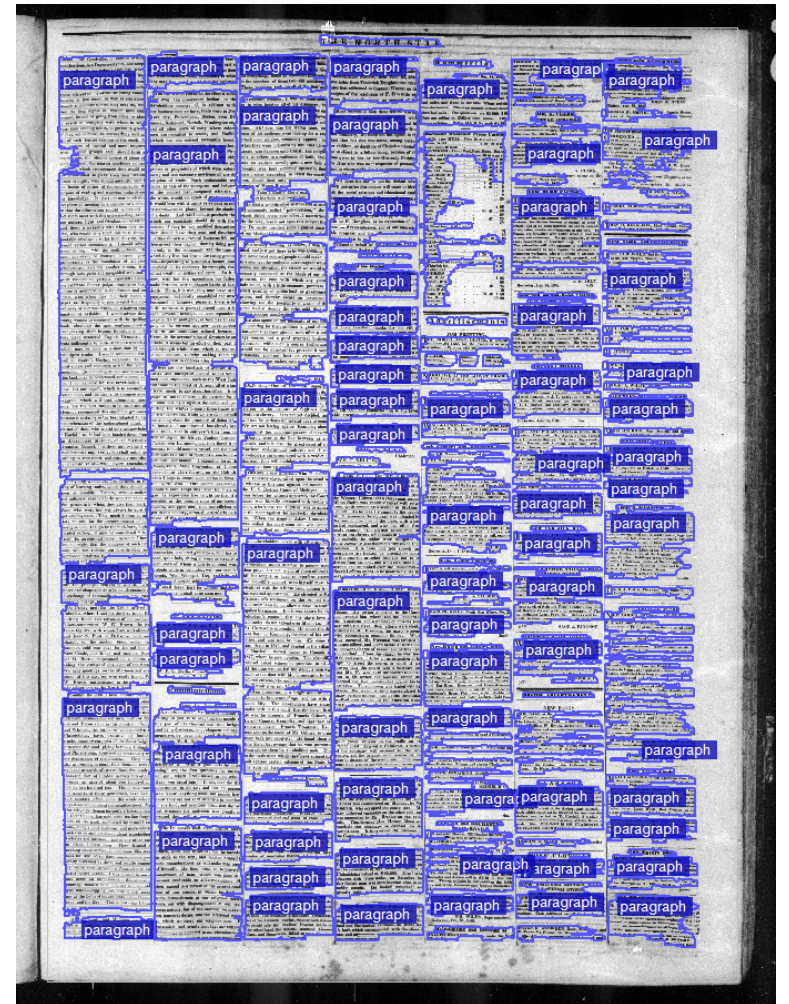
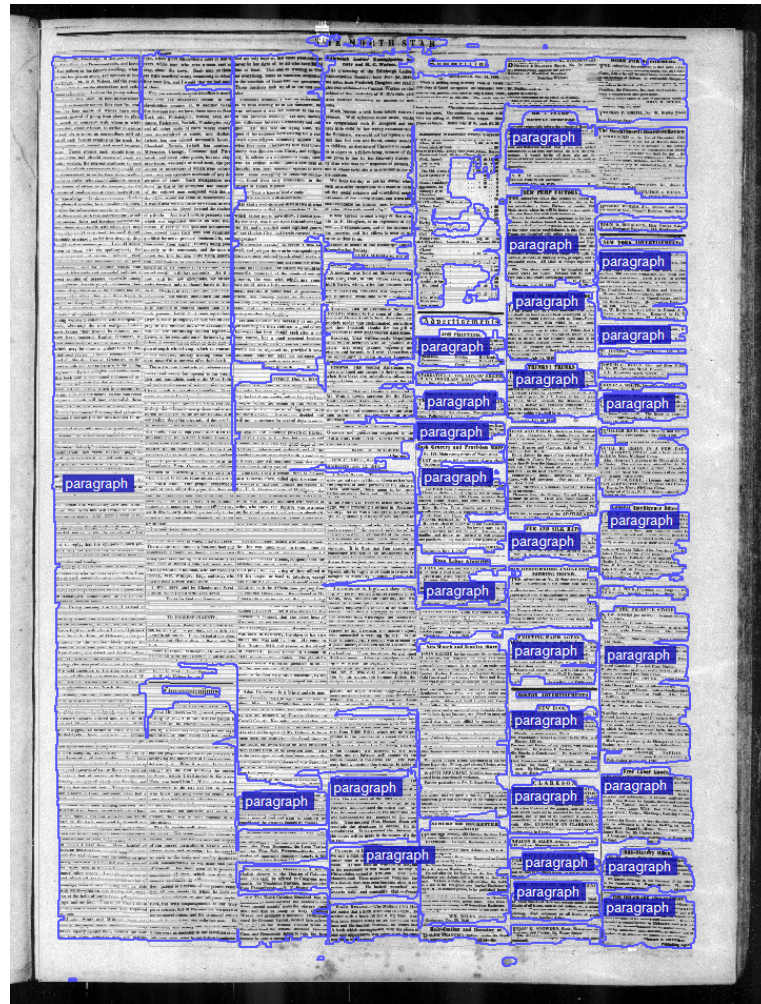
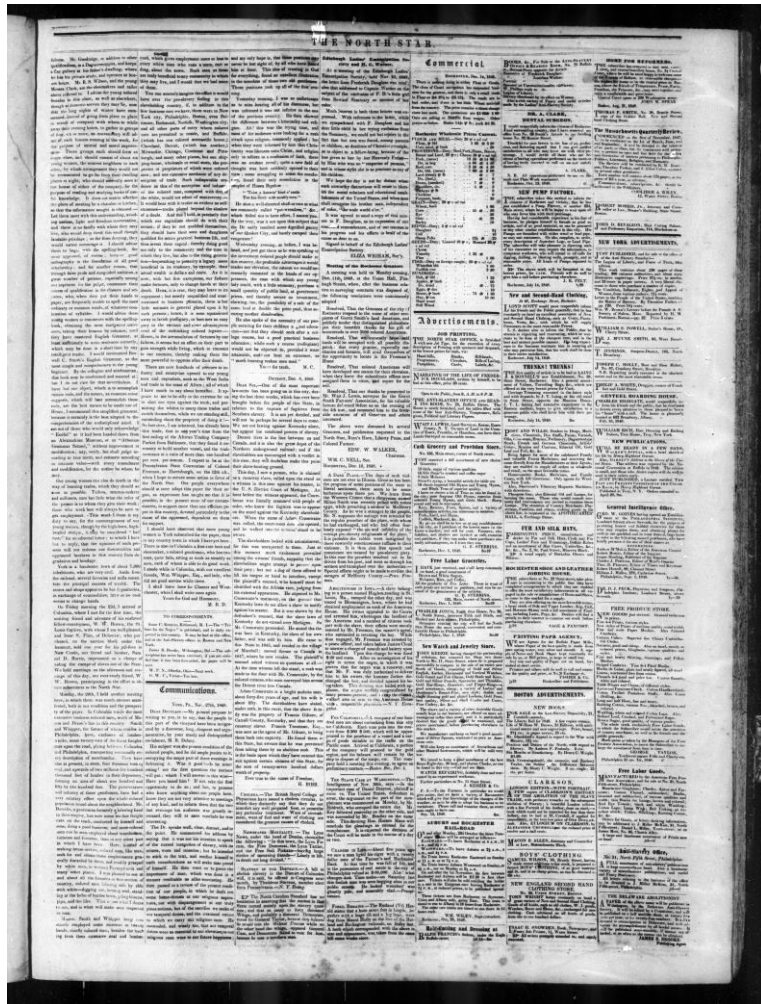
- Fehlerfreie OCR Resultate sind bspw. mit [ocrd_calamari](#) und einem auf dem Datensatz [GT4HistOCR](#) trainierten Modell möglich!
- Ein Vorteil von *Calamari* ggü. *Tesseract*, *OCRopus*, *Kraken*: Voting
- Deep Learning ermöglicht die Erkennung von sowohl Fraktur als auch Antiqua mit einem einzigen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

Layout Analyse

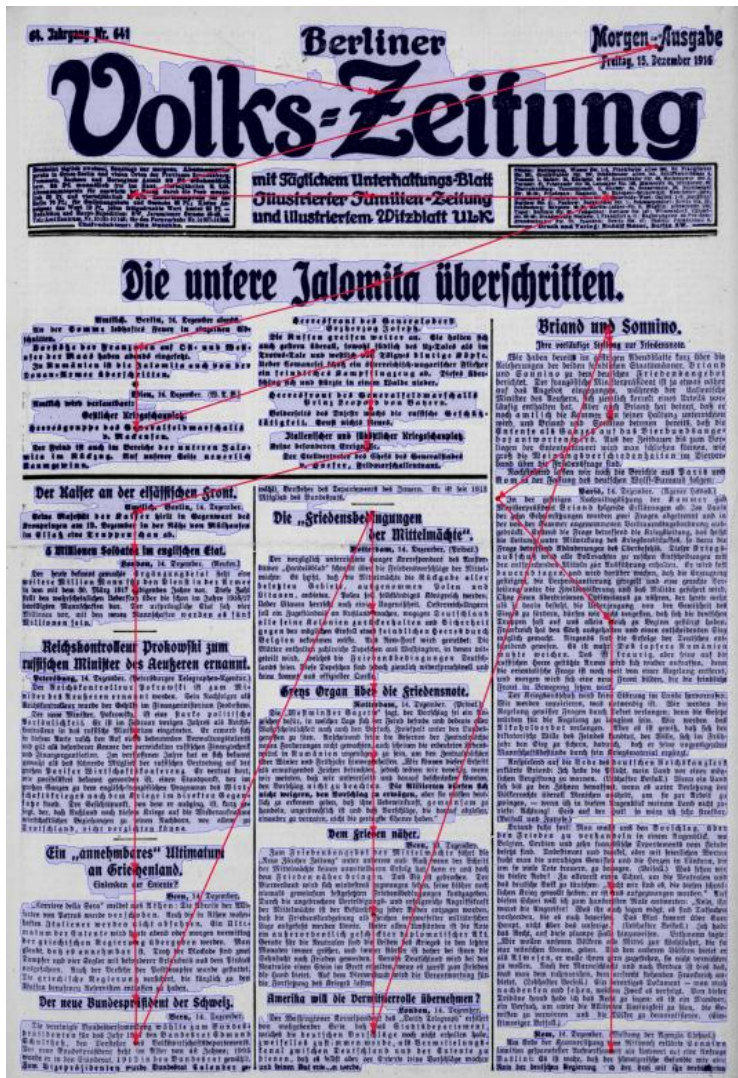
- Trainieren eines Machine Learning Systems (bspw. [dhSegment](#)) basierend auf einer *U-Net* / *ResNet* Architektur für pixel-wise Klassifikation anhand von Ground Truth Daten (mit Augmentation)
 - 1ste Iteration („reines ML“): gute Textzeilensegmentierung, aber Probleme mit Überschriften, Initialen, 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



Herausforderungen



Herausforderungen



Evaluierung

168

JOAN. KEPL. DE STEL. CYGNI.

Ex quibus distantijs extruxerunt Braheani, circumspēctis omnibus locum 16. 18/ Aquarij, latit: 55. 30/ Bor. Hinc invenitur ascensio recta Novæ 300. 46/. declinatio 36. 52/ Borealis. Culminat igitur cum 28. 37/ Capric.

In Hispaniæ parte Andalusia, in Sicilia, Peloponneso, Ionia, Cilicia, Syria, ceterisque locis Terrarum, sub hoc eodem parallelo sitis, per Verticem quotidie transit. Quibus verò est altitudo Poli 53. 8/ , ij horizontem stringit in Septentrione; ut Angliæ, Hollandiæ, Brunsvvigo, Marchiæ, Livoniæ, Moscoviæ. Ulterius versus Septentrionem non occidit.

Species Oloris post accessum Novæ.
N. Novam denotat.



Page number: 168

Header: JOAN. KEPL. DE STEL. CYGNI.

Text 1: Ex quibus distantijs extruxerunt Braheani, circumspēctis omnibus locum 16. 18/ Aquarij, latit: 55. 30/ Bor. Hinc invenitur ascensio recta Novæ 300. 46/. declinatio 36. 52/ Borealis. Culminat igitur cum 28. 37/ Capric.

Text 2: In Hispaniæ parte Andalusia, in Sicilia, Peloponneso, Ionia, Cilicia, Syria, ceterisque locis Terrarum, sub hoc eodem parallelo sitis, per Verticem quotidie transit. Quibus verò est altitudo Poli 53. 8/ , ij horizontem stringit in Septentrione; ut Angliæ, Hollandiæ, Brunsvvigo, Marchiæ, Livoniæ, Moscoviæ. Ulterius versus Septentrionem non occidit.

Text 3: Species Oloris post accessum Novæ.
N. Novam denotat.

Image: A woodcut illustration of the constellation Cygnus (the Swan) with stars marked by crosses.

Evaluierung

168



JOAN. KEPL. DE STEL. CYGNI

Ex quibus distantijs extruxerunt Braheani, circumspēctis omnibus locum 16. 18/ Aquarij, latit: 55. 30/ Bor. Hinc invenitur ascensio recta Novæ 300. 46/. declinatio 36. 52/ Borealis. Culminat igitur cum 28. 37/ Capric.

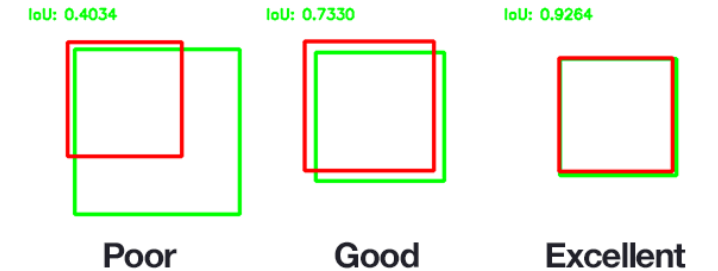
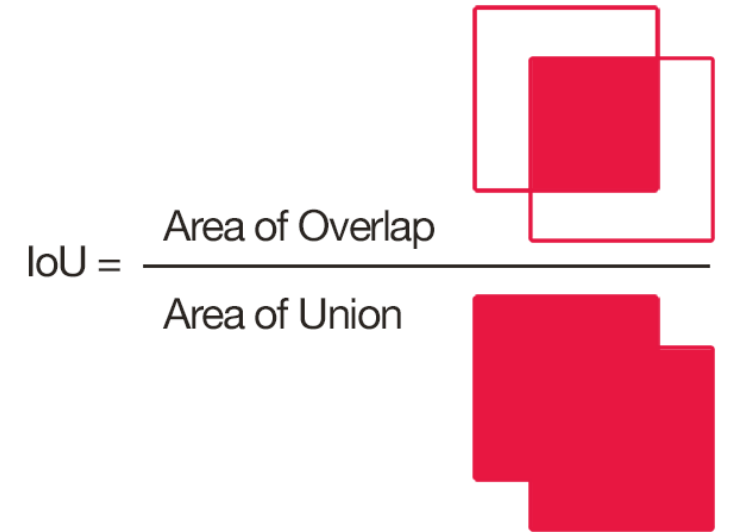
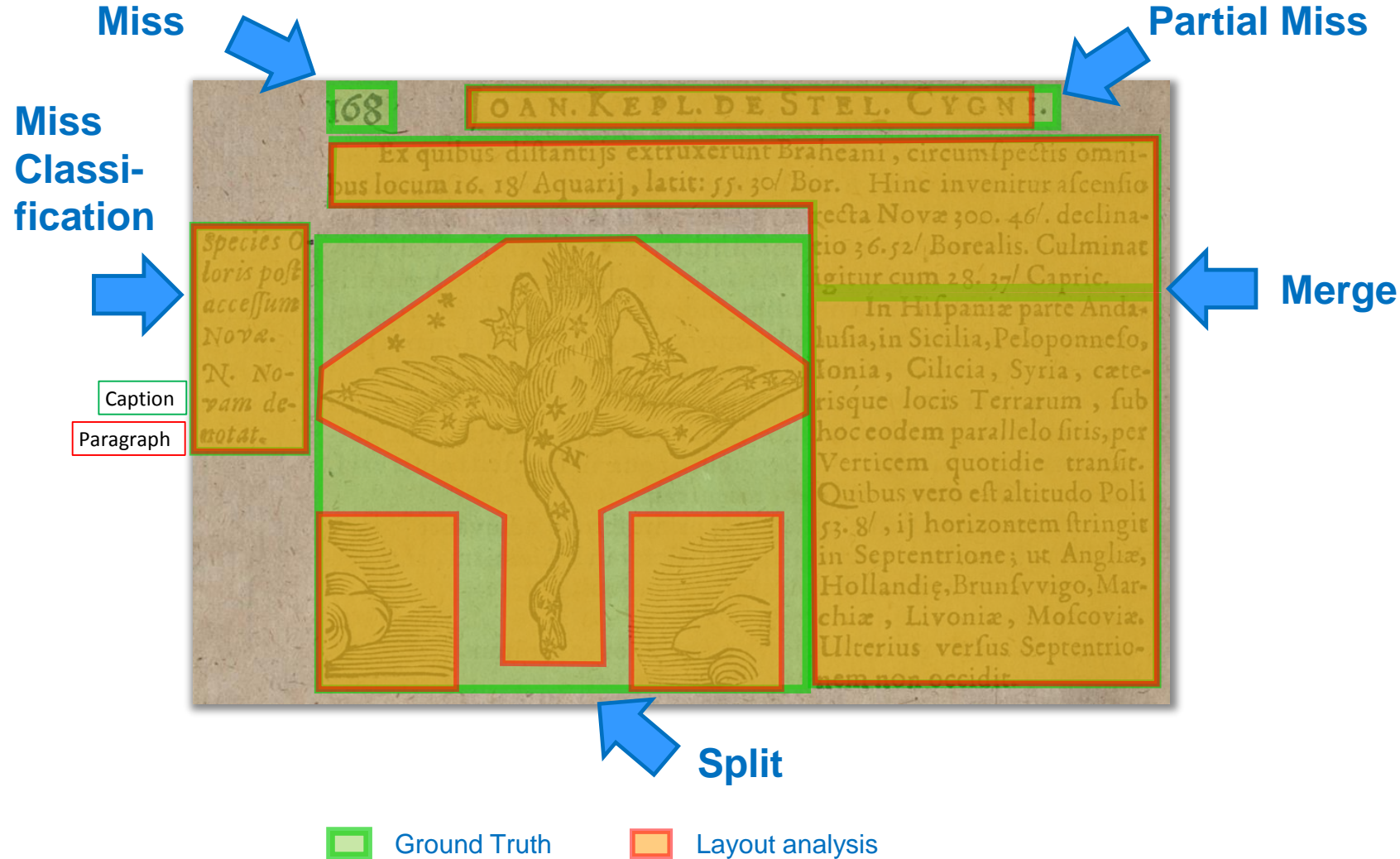
In Hispaniæ parte Andalusia, in Sicilia, Peloponneso, Ionia, Cilicia, Syria, ceterisque locis Terrarum, sub hoc eodem parallelo sitis, per Verticem quotidie transit. Quibus verò est altitudo Poli 53. 8/ , ij horizontem stringit in Septentrione; ut Angliæ, Hollandiæ, Brunsvvigo, Marchiæ, Livoniæ, Moscoviæ. Ulterius versus Septentrionem non occidit.

Species Oloris post accessum Novæ.

N. Novam demotat.



Evaluierung



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