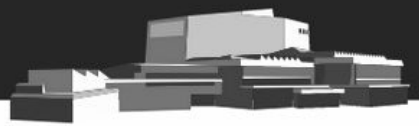


Staatsbibliothek
zu Berlin
Preußischer Kulturbesitz

Mensch.Maschine.Kultur

Welche Chancen und Risiken bringt Künstliche Intelligenz für das kulturelle Erbe?

Clemens Neudecker | Staatsbibliothek zu Berlin - Preußischer Kulturbesitz

HAW Hamburg Ringvorlesung “Bibs & Bits” | 9. Januar 2024

Staatsbibliothek zu Berlin - Preußischer Kulturbesitz

- gehört zur Stiftung Preußischer Kulturbesitz (SPK)
- ist an zwei Standorten in Berlin an 7 Tagen die Woche kostenlos für die Benutzung geöffnet
- sammelt seit 1661 wissenschaftlich relevante Literatur aus allen Sprachen, Zeiten und Ländern
- besitzt Hauptbestand mit ca. 12 Mio. Büchern, jährlicher Zuwachs von etwa 100,000 Titeln
- **Digitalisierte Sammlungen** geben Zugang zu über 210,000 digitalisierten Dokumenten, mit Public Domain Lizenz und via IIIF API
- **Stabi Lab** für Experimente, Datensets, Digital Humanities

Staatsbibliothek zu Berlin
Preußischer Kulturbesitz

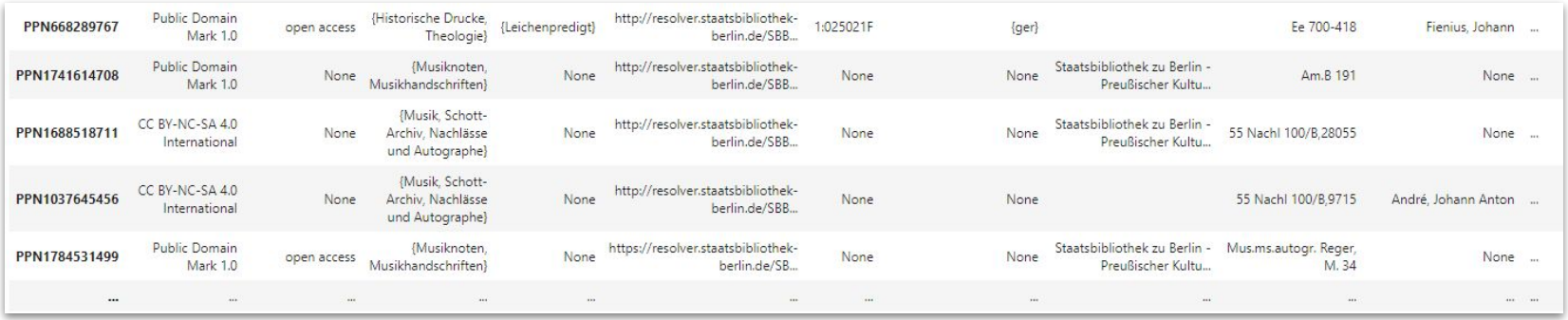
KI (bzw. Machine Learning) in der Stabi Berlin

- Drittmittelprojekte
  - 2016—2024: **OCR-D** (DFG)
  - 2018—2021: **Qurator** (BMBF)
  - 2022—2025: **Mensch.Maschine.Kultur** (BKM)
- Team
  - 6x 100% Machine Learning Engineer
  - 2x 50% Bibliothekswissenschaftler:in
  - 1x 100% Forschungsdatenmanager:in
- Hardware
  - GPU Server mit 4x NVIDIA Tesla GPUs V100/A100
- Ziele
  - Erforschung und Entwicklung von Open Source Technologien für historische Kulturdaten
  - Durchsuchbarkeit (Texte und Bilder) sowie Erschließung und Strukturierung der Inhalte
  - Bereitstellung als offene und maschinenlesbare Daten (“Collections as Data”)
  - Ethisch, rechtlich, sozial verantwortungsvoller Umgang mit problematischen Inhalten und KI

Datenanalyse und Statistik

- Extraktion und Aufbereitung von diversen Metadaten (METS, MODS, ALTO) für statistische Analysen mit Python Pandas

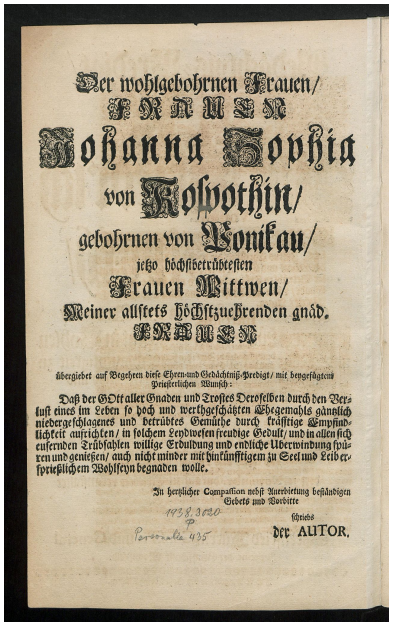
  https://github.com/qurator-spk/mods4pandas

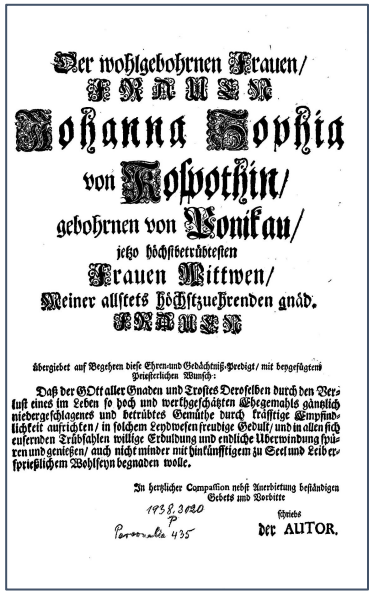


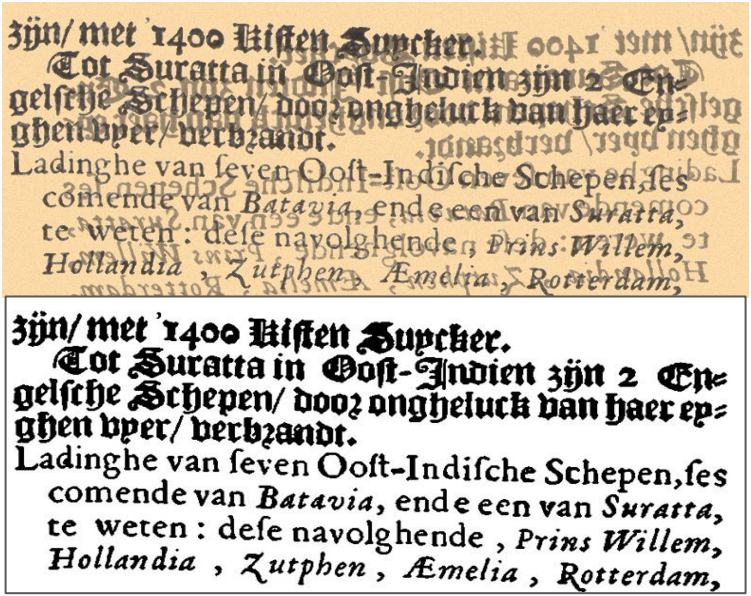
PPN668289767	Public Domain Mark 1.0	open access	{Historische Drucke, Theologie}	{Leichenpredigt}	http://resolver.staatsbibliothek-berlin.de/SBB...	1:025021F	{ger}		Ee 700-418	Fienius, Johann	...
PPN1741614708	Public Domain Mark 1.0	None	{Musiknoten, Musikhandschriften}	None	http://resolver.staatsbibliothek-berlin.de/SBB...	None	None	Staatsbibliothek zu Berlin - Preußischer Kultu...	Am.B 191	None	...
PPN1688518711	CC BY-NC-SA 4.0 International	None	{Musik, Schott-Archiv, Nachlässe und Autographe}	None	http://resolver.staatsbibliothek-berlin.de/SBB...	None	None	Staatsbibliothek zu Berlin - Preußischer Kultu...	55 Nachl 100/B,28055	None	...
PPN1037645456	CC BY-NC-SA 4.0 International	None	{Musik, Schott-Archiv, Nachlässe und Autographe}	None	http://resolver.staatsbibliothek-berlin.de/SBB...	None	None		55 Nachl 100/B,9715	André, Johann Anton	...
PPN1784531499	Public Domain Mark 1.0	open access	{Musiknoten, Musikhandschriften}	None	https://resolver.staatsbibliothek-berlin.de/SB...	None	None	Staatsbibliothek zu Berlin - Preußischer Kultu...	Mus.ms.autogr. Reger, M. 34	None	...
...	...	...	...	...	...	...	...	...	...	...	...

Bildoptimierung und Binarisierung

- *A hybrid CNN-Transformer model for Historical Document Image Binarization*, 2023.
  https://doi.org/10.1145/3604951.3605508 | **https://github.com/qurator-spk/sbb_binarization**

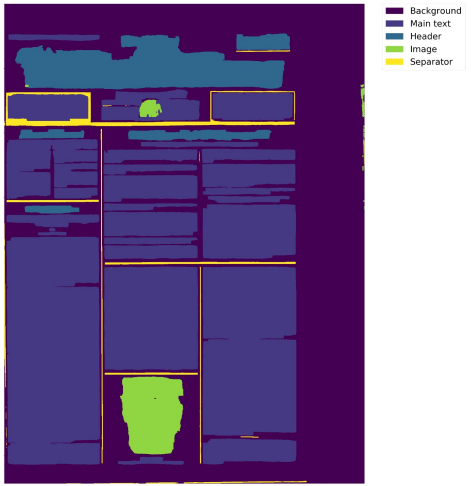






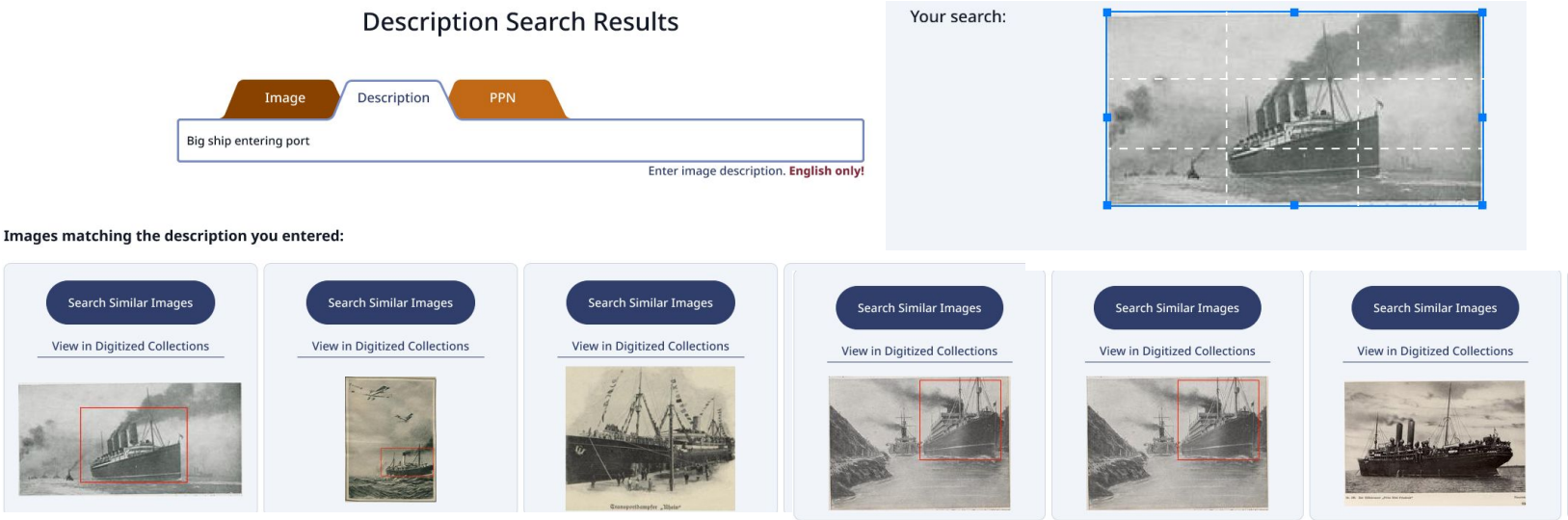
Layout- und Strukturanalyse

- *Document Layout Analysis with Deep Learning and Heuristics*, 2023.
  https://doi.org/10.1145/3604951.3605513 | **https://github.com/qurator-spk/eynollah**



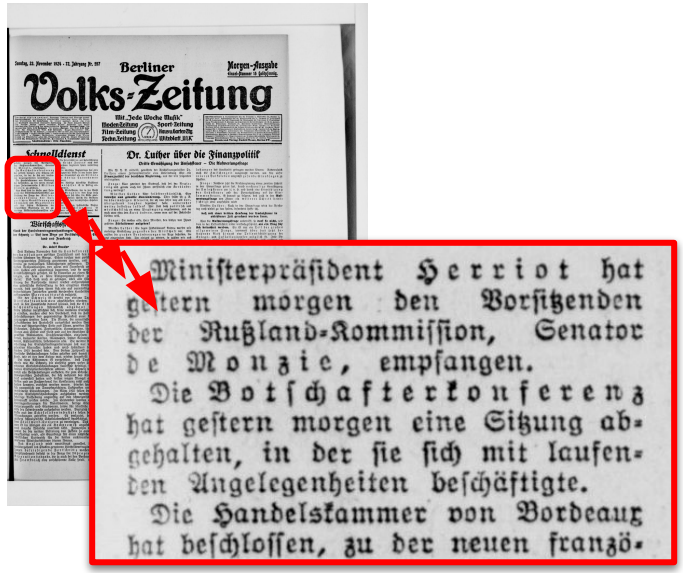
Bildähnlichkeitssuche und Text-Bild-Suche

- *Gauging the Limitations of Natural Language Supervised Text-Image Metrics Learning by Iconclass Visual Concepts*, 2023.
  https://doi.org/10.1145/3604951.3605516 | **https://github.com/qurator-spk/sbb_images**



Texterkennung (OCR)

- *OCR-D: An end-to-end open source OCR framework for historical printed documents*, 2019. **https://doi.org/10.1145/3322905.3322917** | **https://github.com/OCR-D/core**

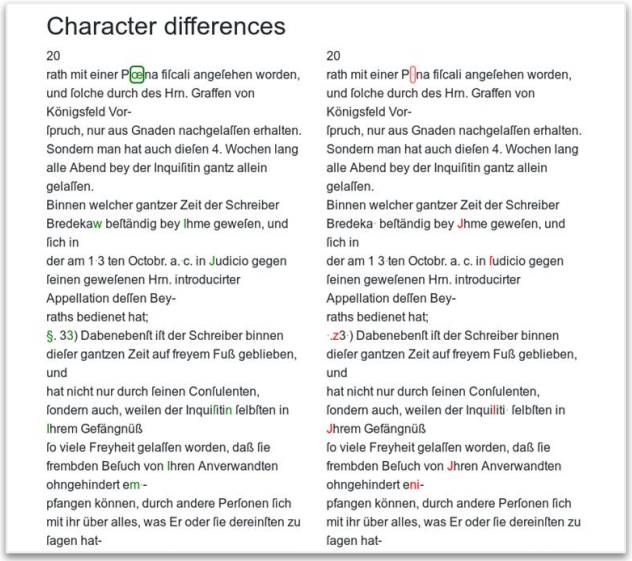


Ministerpräsident Hcrri 0 t hat
Die Feierlichkeiten zur UcbcrfUhrung
geilern morgen de» Vorsitzenden
der Autzlanü-Komniission, Cenator
dl M 0 ll z i e, empfangen.
Die Botschafterkonferenz
bat gestern morgen eine Citzung ab,
gehalten, in der sie sich mit laufen-
ten Angelegenheiten beschäftigte.
Die Handelskammer von Bordeaux
bat defchloffen, zu der neuen franzö_
sischen Inlandsanleihe 1 Million
Francs zu zeichne».
Aus Genf sind in Sofia zwei Dele-
gierte der Bälkerdundskommifsion zur
Prüfung der Frage der Massen,
auswanderung der bul.
garifchen Bevölkerung aus
Thrazien und Mazedonien
und der Ictzien Beschwerde der
bulgarischen Regierung an die zu
ständige Dölkerbundskommisston ein-
getroffen.

Miniſterpräſident Herriot hat
geſtern morgen den Vorſitzenden
der Rußland- Kommiſſion, Senator
de Monzie, empfangen.
Die Botſchafterkonferenz
hat geſtern morgen eine Sitzung ab–
gehalten, in der ſie ſich mit laufen-
den Angelegenheiten beſchäftigte.
Die Handelskammer von Bordeaux
hat beſchloſſen, zu der neuen franzö-
ſiſchen Jnlandsanleihe 1 Million
Francs zu zeichnen.
Aus Genf ſind in Sofia zwei Dele-
gierte der Völkerbundskommiſſion zur
Prüfung der Frage der Maſſen-
auswanderung der bul-
gariſchen Bevölkerung aus
Thrazien und Mazedonien
und der letzten Beſchwerde der
bulgariſchen Regierung an die zu–
ſtändige Völkerbundskommiſſion ein-
getroffen.

OCR Evaluation und Qualitätssicherung

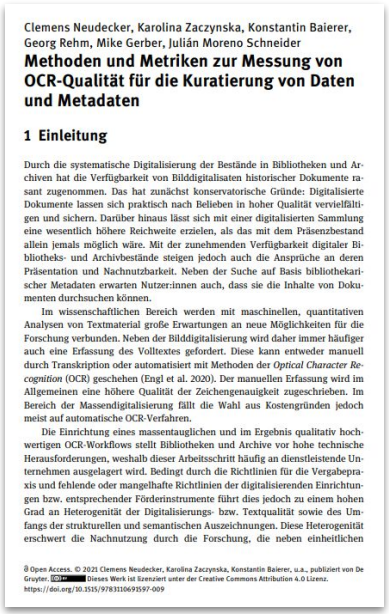
- Verlässliche Ermittlung der OCR Qualität (CER, WER, Reading Order) mit Ground Truth **https://github.com/qurator-spk/dinglehopper**



Character differences

20
rath mit einer Pcena fiſcali angeſehen worden,
und ſolche durch des Hrn. Graffen von
Königsfeld Vor-
ſpruch, nur aus Gnaden nachgelaſſen erhalten.
Sondern man hat auch dieſen 4. Wochen lang
alle Abend bey der Inquiſition gantz allein
gelaſſen.
Binnen welcher gantzer Zeit der Schreiber
Bredekaw beſtändig bey Ihme geweſen, und
ſich in
der am 1 3 ten Octobr. a. c. in Judicio gegen
ſeinen geweſenen Hrn. introducirter
Appellation deſſen Bey-
raths bedienet hat;
§. 33) Dabenebenſt iſt der Schreiber binnen
dieſer gantzen Zeit auf freyem Fuß geblieben,
und
hat nicht nur durch ſeinen Confulenten,
ſondern auch, weilen der Inquiſitin ſelbſten in
Ihrem Gefängnüß
ſo viele Freyheit gelaſſen worden, daß ſie
frembden Beſuch von Ihren Anverwandten
ohngehindert em -
pfangen können, durch andere Perſonen ſich
mit ihr über alles, was Er oder ſie dereinſten zu
ſagen hat-

20
rath mit einer P{na fiſcali angeſehen worden,
und ſolche durch des Hrn. Graffen von
Königsfeld Vor-
ſpruch, nur aus Gnaden nachgelaſſen erhalten.
Sondern man hat auch dieſen 4. Wochen lang
alle Abend bey der Inquiſition gantz allein
gelaſſen.
Binnen welcher gantzer Zeit der Schreiber
Bredeka beſtändig bey Jhme geweſen, und
ſich in
der am 1 3 ten Octobr. a. c. in ſudicio gegen
ſeinen geweſenen Hrn. introducirter
Appellation deſſen Bey-
raths bedienet hat;
.z3) Dabenebenſt iſt der Schreiber binnen
dieſer gantzen Zeit auf freyem Fuß geblieben,
und
hat nicht nur durch ſeinen Confulenten,
ſondern auch, weilen der Inquiliti ſelbſten in
Jhrem Gefängnüß
ſo viele Freyheit gelaſſen worden, daß ſie
frembden Beſuch von Jhren Anverwandten
ohngehindert eni-
pfangen können, durch andere Perſonen ſich
mit ihr über alles, was Er oder ſie dereinſten zu
ſagen hat-



Clemens Neudecker, Karolina Zaczynska, Konstantin Baierer, Georg Rehm, Mike Gerber, Julián Moreno Schneider

Methoden und Metriken zur Messung von OCR-Qualität für die Kuratierung von Daten und Metadaten

1 Einleitung

Durch die systematische Digitalisierung der Bestände in Bibliotheken und Archiven hat die Verfügbarkeit von Bilddigitalisaten historischer Dokumente rasant zugenommen. Das hat zunächst konservatorische Gründe: Digitalisierte Dokumente lassen sich praktisch nach Belieben in hoher Qualität vervielfältigen und sichern. Darüber hinaus lässt sich mit einer digitalisierten Sammlung eine wesentlich höhere Reichweite erzielen, als das mit dem Präsenzbestand allein jemals möglich wäre. Mit der zunehmenden Verfügbarkeit digitaler Bibliotheks- und Archivbestände steigen jedoch auch die Ansprüche an deren Präsentation und Nachnutzbarkeit. Neben der Suche auf Basis bibliothekarischer Metadaten erwarten Nutzer:innen auch, dass sie die Inhalte von Dokumenten durchsuchen können.

Im wissenschaftlichen Bereich werden mit maschinellen, quantitativen Analysen von Textmaterial große Erwartungen an neue Möglichkeiten für die Forschung verbunden. Neben der Bilddigitalisierung wird daher immer häufiger auch eine Erfassung des Volltextes gefordert. Diese kann entweder manuell durch Transkription oder automatisiert mit Methoden der *Optical Character Recognition* (OCR) geschehen (Engl et al. 2020). Der manuellen Erfassung wird im Allgemeinen eine höhere Qualität der Zeichengenauigkeit zugeschrieben. Im Bereich der Massendigitalisierung fällt die Wahl aus Kostengründen jedoch meist auf automatische OCR-Verfahren.

Die Einrichtung eines massentauglichen und im Ergebnis qualitativ hochwertigen OCR-Workflows stellt Bibliotheken und Archive vor hohe technische Herausforderungen, weshalb dieser Arbeitsschritt häufig an dienstleistende Unternehmen ausgelagert wird. Bedingt durch die Richtlinien für die Vergabepraxis und fehlende oder mangelhafte Richtlinien der digitalisierenden Einrichtungen bzw. entsprechender Förderinstrumente führt dies jedoch zu einem hohen Grad an Heterogenität der Digitalisierungs- bzw. Textqualität sowie des Umfangs der strukturellen und semantischen Auszeichnungen. Diese Heterogenität erschwert die Nachnutzung durch die Forschung, die neben einheitlichen

https://doi.org/10.1515/9783110693597-009

A survey of OCR evaluation tools and metrics

Clemens Neudecker
Konstantin Baierer
Mike Gerber
www.staatsbibliothek-berlin.de
Staatsbibliothek zu Berlin - Preußischer Kulturbesitz
Berlin, Germany

Christian Clausner
Apostolos Antonacopoulos
Stefan Pletschacher
www.primaresearch.org
Pattern Recognition and Image Analysis Lab (PRImA)
University of Salford
Greater Manchester, United Kingdom

ABSTRACT

The millions of pages of historical documents that are digitized in libraries are increasingly used in contexts that have more specific requirements for OCR quality than keyword search. How to comprehensively, efficiently and reliably assess the quality of OCR results against the background of mass digitization, when ground truth can only ever be produced for very small numbers? Due to gaps in specifications, results from OCR evaluation tools can return different results, and due to differences in implementation, even commonly used error rates are often not directly comparable. OCR evaluation metrics and sampling methods are also not sufficient where they do not take into account the accuracy of layout analysis, since for advanced use cases like Natural Language Processing or the Digital Humanities, accurate layout analysis and detection of the reading order are crucial. We provide an overview of OCR evaluation metrics and tools, describe two advanced use cases for OCR results, and perform an OCR evaluation experiment with multiple evaluation tools and different metrics for two distinct datasets. We analyze the differences and commonalities in light of the presented use cases and suggest areas for future work.

CCS CONCEPTS

• **Applied computing** → **Optical character recognition**; *Document analysis*; *Graphics recognition and interpretation*; • **Information systems** → **Digital libraries and archives**.

KEYWORDS

optical character recognition, evaluation, accuracy, metrics

ACM Reference Format:
Clemens Neudecker, Konstantin Baierer, Mike Gerber, Christian Clausner, Apostolos Antonacopoulos, and Stefan Pletschacher. 2021. A survey of OCR evaluation tools and metrics. In *The 6th International Workshop on Historical Document Imaging and Processing (HIP '21), September 5–6, 2021, Lausanne, Switzerland.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3476887.3476888

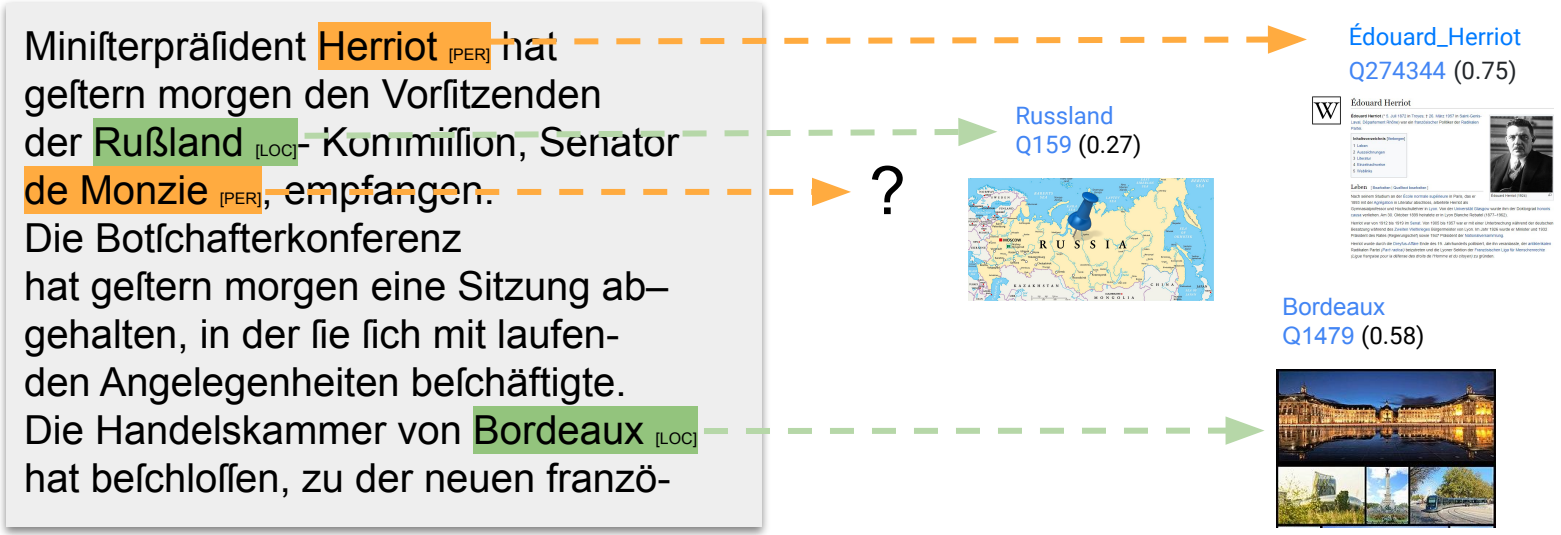
HIP'21, September 05–06, 2021, Lausanne, Switzerland

ACM ISBN 978-1-4503-8690-6/21/09…$15.00
https://doi.org/10.1145/3476887.3476888

[1] Google estimated in 2010 that there are around 130M unique books published (cf. http://booksearch.blogspot.com/2010/08/books-of-world-stand-up-and-be-counted.html) and digitized 40M in October 2019: https://www.blog.google/products/search/15-years-google-books/.
[2] https://ocr-d.de/en/gt-guidelines/trans/ocr_d_koordinatensgrenzen_codierung.html
[3] Medieval Unicode Font Initiative. https://folk.uib.no/hnooh/mufi/.
[4] Unicode Standard, Chapter 16: Special Areas and Format Characters.

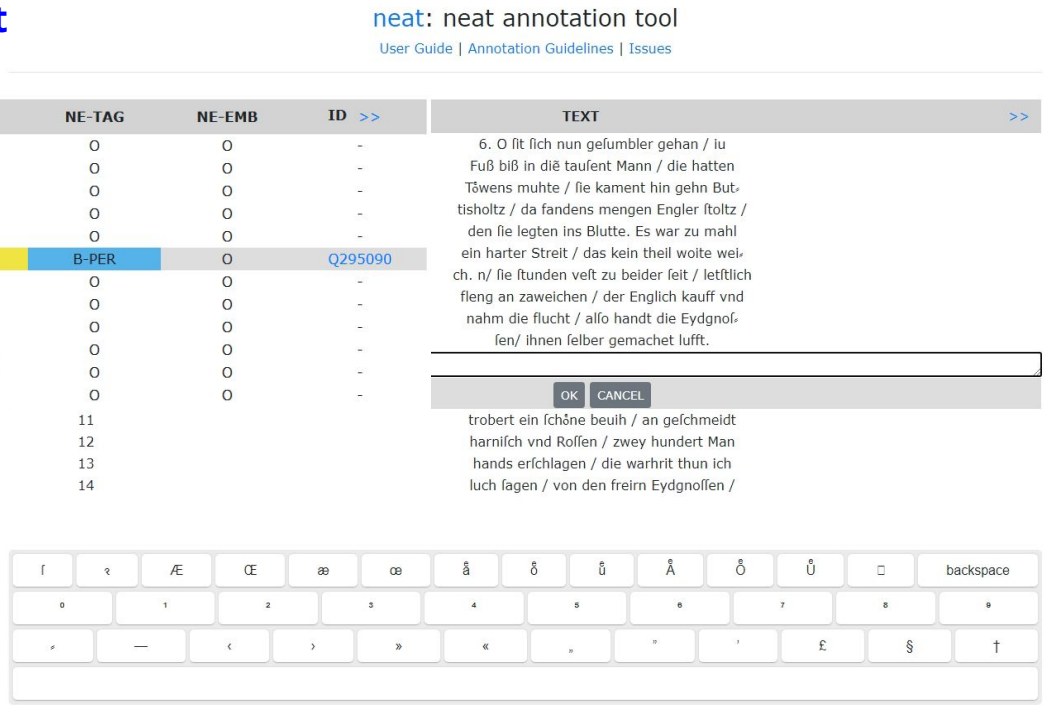
Named Entity Recognition, Entity Linking

- *BERT for Named Entity Recognition in Contemporary and Historic German*, 2019.
  https://konvens.org/proceedings/2019/papers/KONVENS2019_paper_4.pdf |
  https://github.com/qurator-spk/sbb_ner | **https://github.com/qurator-spk/sbb_ned**



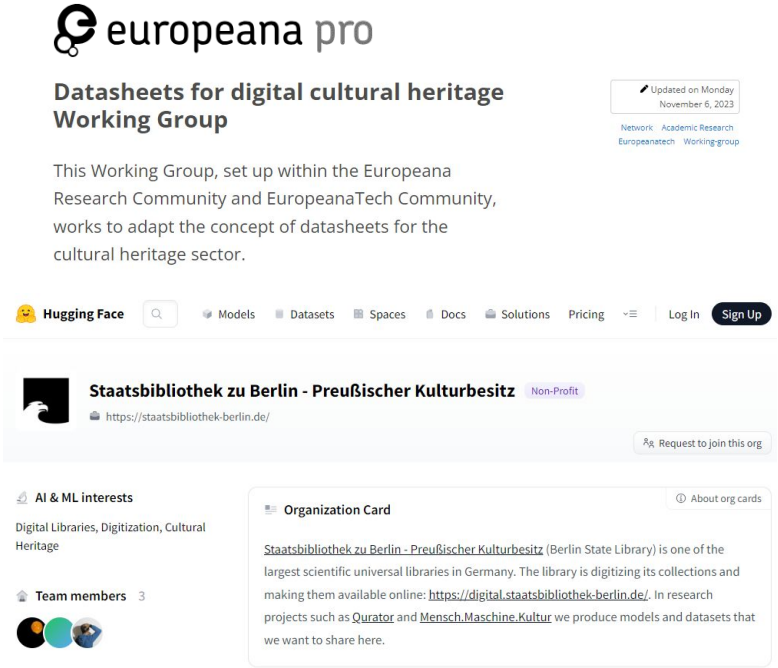
Transkription und Annotation

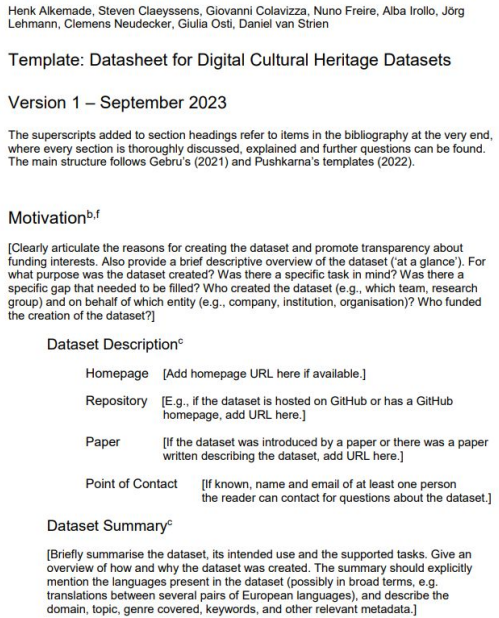
- Annotation von Named Entities, Korrektur und Transkription von Volltexten, Auszeichnung von Bilddaten und Elementen direkt im Browser
  https://github.com/qurator-spk/neat



Datenbereitstellung und Kuratierung

- *Datasheets for Digital Cultural Heritage Datasets*, 2023.
  https://doi.org/10.5334/johd.124 | **https://doi.org/10.5281/zenodo.8375034**





Was ist überhaupt KI?

- Der Begriff “Künstliche Intelligenz“ wird als problematisch angesehen und sollte besser vermieden werden, da er einen Anthropomorphismus darstellt - als ob es sich hierbei um eine dem Menschen vergleichbare Intelligenz handelt. Dies ist aber eher nicht der Fall.
- Besser wäre entweder allgemein von “maschinellem Lernen” bzw. „deep learning“ zu sprechen oder spezifischer von “stochastischen Vorhersagemodellen” (provokanter „stochastic parrots“, vgl. **Bender et al. 2021**)
- Grundsätzlich gilt: aus möglichst vielen repräsentativen Ausgangsdaten (Beispielen) werden Wahrscheinlichkeitsmodelle trainiert, um diese dann auf weitere Daten anwenden zu können.
- Die Qualität eines Modells (der “KI”) hängt also maßgeblich davon ab, wie umfangreich, qualitativ hochwertig und vielfältig die zum Training verwendeten Ausgangsdaten sind.
- Für die aktuellen KI-Modelle sind jedoch die Antworten auf viele relevante Fragen unbekannt:
    - Welche Daten wurden zum Training verwendet und wo können diese eingesehen werden?
    - Welche Qualitäts- und sonstigen Auswahlkriterien kamen dabei zur Anwendung?
    - Welche Personen waren ggf. an der Annotation von Daten beteiligt?
    - Welche Kontaktmöglichkeiten gibt es um Probleme und Fehler zu melden?

Herausforderungen bei Verwendung von KI

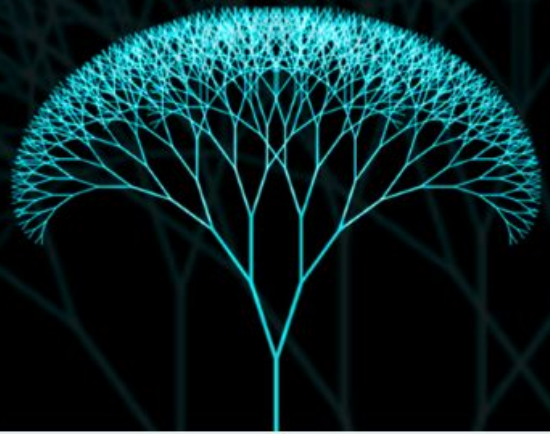
- Aufgrund von geltendem Urheberrecht werden vor allem historische Werke digitalisiert – aber diese enthalten historische Begriffe und Rechtschreibung, für die eine KI erst angepasst bzw. trainiert werden muss
- Digitalisate und Metadaten liegen nicht in der geeigneten Form und in den Formaten vor, wie sie für das Training von KI benötigt werden und müssen daher zunächst evaluiert, konvertiert und/oder (von Expert:innen) transkribiert und annotiert werden
- Kulturelle Kontexte (z.B. Zeit, Ort) müssen bei der Verwendung von Kulturerbe und KI Berücksichtigung finden
- Ethisch, sozial oder rechtlich problematische Inhalte (wie z.B. Kolonialismus, Nationalsozialismus oder die Unterrepräsentation von marginalisierten Gruppen) in den Daten müssen identifiziert und entsprechend vorsichtig behandelt und kontextualisiert werden
- Spezielle Hardware (GPUs) und Expertise (Einwerbung von Personal mit entsprechenden Kompetenzen) wird benötigt

Potentiale von Kulturerbeeinrichtungen und KI

- Der Einsatz von KI bietet viele Möglichkeiten zur effizienten Erschließung, Analyse und Anreicherung von Digitalisaten sowie für neue Services für Nutzende und die Wissenschaft
- Durch die fortschreitende Massendigitalisierung verfügen Kulturerbeeinrichtungen über große und stetig wachsende Mengen an (zumeist offenen) Daten für Training und Verbesserung von KI
- In den jeweiligen Fachbereichen in Kulturerbeeinrichtungen gibt es große Expertise zu den Sammlungen und Inhalten, von denen eine KI lernen bzw. profitieren kann
- Grundsätzlich besteht ein (vor allem im Vergleich zu großen Tech-Unternehmen) hohes Qualitätsbewusstsein und Sensibilität bei der Erstellung, Pflege und Nutzung von Daten
- Als öffentliche Einrichtungen und Dienstleister für Forschung und Wissenschaft werden Transparenz, Datenschutz und Verantwortung im Umgang mit Daten und KI ernst genommen – auch über sehr lange Zeiträume

“In closing, we advocate for a turn in the culture towards **carefully collected** datasets, **rooted in their original contexts**, distributed only in ways that **respect the intellectual property and privacy rights** of data creators and data subjects, and constructed **in conversation with the relevant scientific and scholarly fields** required to create datasets that faithfully model tasks and tasks which target **relevant and realistic capabilities**. Such datasets will undoubtedly be **more expensive to create, in time, money and effort**, and therefore smaller than today’s most celebrated benchmarks. This, in turn, will encourage work on approaches to machine learning (and to artificial intelligence beyond machine learning) that **go beyond the current paradigm of techniques idolizing scale**. Should this come to pass, we predict that machine learning as a field will be better positioned to **understand how its technology impacts people** and to design solutions that work with **fidelity and equity** in their deployment contexts.”

A. Paullada, I. D. Raji, E. M. Bender, E. Denton, A. Hanna. Data and its (dis)contents: A survey of dataset development and use in machine learning research, 2021. **https://doi.org/10.1016/j.patter.2021.100336**



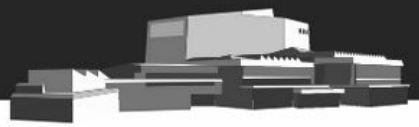
AI4LAM

Artificial Intelligence for Libraries, Archives & Museums

Cultural AI
a lab for
culturally
valued AI

Cultural AI is the study, design and development of socio-technological AI systems that are implicitly or explicitly aware of the subtle and subjective complexity of human culture.


Culture for AI
AI for Culture



Staatsbibliothek
zu Berlin
Preußischer Kulturbesitz

Danke für die Aufmerksamkeit!

Fragen?

Clemens Neudecker | Staatsbibliothek zu Berlin - Preußischer Kulturbesitz

HAW Hamburg Ringvorlesung “Bibs & Bits” | 9. Januar 2024