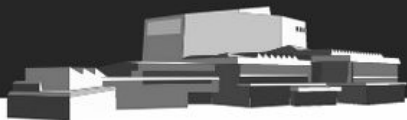




**Staatsbibliothek
zu Berlin**
Preußischer Kulturbesitz



Künstliche Intelligenz für Kulturdaten und die Kultur(en) der KI

Welche Rollen spielen die Bibliotheken?

Clemens Neudecker | Staatsbibliothek zu Berlin - Preußischer Kulturbesitz
Niedersächsischer Bibliothekstag | 8. November 2024

I. Eine kurze Einführung in KI

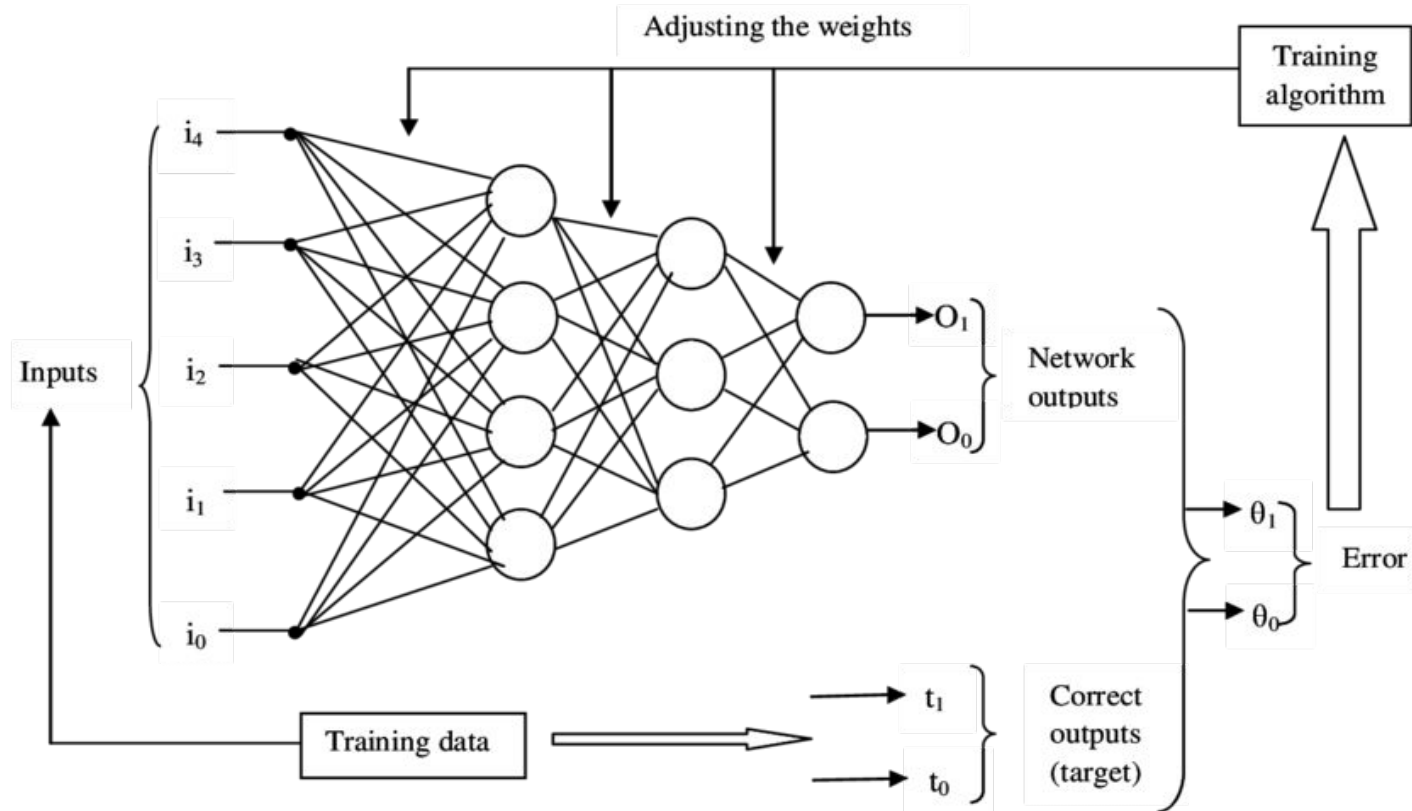
Was genau ist Künstliche Intelligenz (bzw. Deep Learning)?

- Der Begriff “Künstliche Intelligenz” wird in Expertenkreisen als durchaus 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 nicht der Fall.
- Wir bevorzugen es 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 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 aktuell verbreitete KI Modelle fehlen uns jedoch die Antworten auf viele wichtige Fragen:
 - Welche Daten sind für das Training verwendet worden?
 - Welche Qualitäts- und Auswahlkriterien wurden bei dabei angewendet?
 - Welche Personen und Richtlinien wurden für die Annotation von Daten eingesetzt?
 - Wohin kann man sich bei Fehlern und Problemen wenden?

Beispiele

- Um das „Trainieren“ einer KI / eines Modells zu veranschaulichen können zwei einfache Beispiele dienen:
- Bildverarbeitung (Computer Vision)
 - Eine KI / ein Modell soll trainiert werden um Äpfel von Orangen zu unterscheiden. Der KI werden dazu so lange verschiedene Bilder von Äpfeln und Orangen gezeigt, bis die KI / das Modell für ein noch nicht gesehenes Bild selbst korrekt entscheiden kann, ob es sich dabei um einen Apfel oder eine Orange handelt.
- Textverarbeitung mit Sprachmodellen wie z.B. ChatGPT (Natural Language Processing)
 - Eine KI / ein Modell soll trainiert werden um Texte zu bestimmten Themen zu verfassen oder Antworten auf Fragen zu geben. Die KI bekommt dazu sehr viele Texte gezeigt, in denen einzelne Wörter „maskiert“ bzw. ausgeblendet werden, also etwa: „Menschen gehen gerne in Bibliotheken um dort [MASK] zu lesen“.
 - Für die Frage: „Warum gehen Menschen gerne in Bibliotheken?“ macht das Modell dann eine Vorhersage, welches Wort - basierend auf den Trainingsdaten - an Stelle von [MASK] am wahrscheinlichsten stehen könnte, hier also z.B. „Menschen gehen gerne in Bibliotheken um dort Bücher zu lesen.“
- Seit ca. 2 Jahren werden die beiden Modalitäten Bild und Text in sogenannten multimodalen Modellen zusammengeführt. Dies ist möglich, da ein KI Modell intern alle Daten und Informationen nur noch als Zahlen (Parameter und Gewichte) enthält auf denen dann Berechnungen durchgeführt werden können.

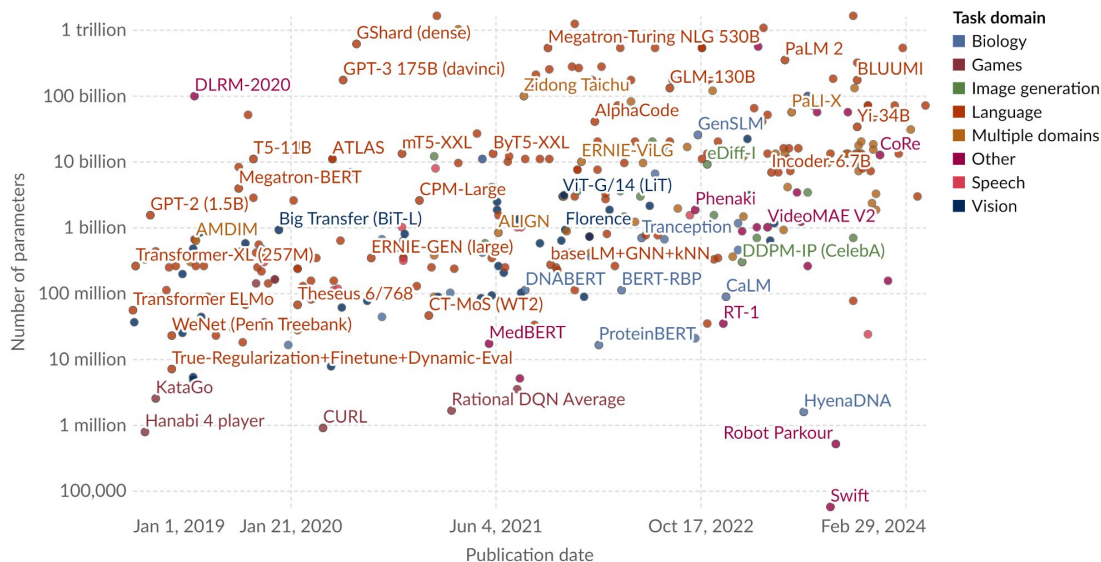
Wie lernt ein neuronales Netz / wie wird eine KI trainiert?



Die Macht der Größe und der Hunger nach Energie

Parameters in notable artificial intelligence systems

Parameters are variables in an AI system whose values are adjusted during training to establish how input data gets transformed into the desired output; for example, the connection weights in an artificial neural network.



Data source: Epoch (2024)

OurWorldInData.org/artificial-intelligence | CC BY

Our World in Data

Consumption	CO ₂ e (lbs)
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

Quelle: Emma Strubell, Ananya Ganesh, Andrew McCallum (2019). Energy and Policy Considerations for Deep Learning in NLP.

II. Maschinelles Lernen in der Stabi Berlin

Staatsbibliothek zu Berlin

- gehört zur Stiftung Preußischer Kulturbesitz (SPK)
- ist an 7 Tagen in zwei Standorten für die Benutzung geöffnet
- sammelt seit 1661 wissenschaftliche Literatur in allen Sprachen
- Bestand von ca. 12 Mio. Objekten mit etwa 100,000 Titeln jährlichem Zuwachs
- **Digitalisierte Sammlungen** bieten Zugang zu >220,000 digitalisierten Dokumenten mit Public Domain Lizenz und IIF Schnittstelle
- verwaltet aktuell rund 7,5 PetaBytes (7500 TeraBytes) an digitalen Ressourcen
- **Stabi Lab** für Experimente, Veranstaltungen, Digital Humanities

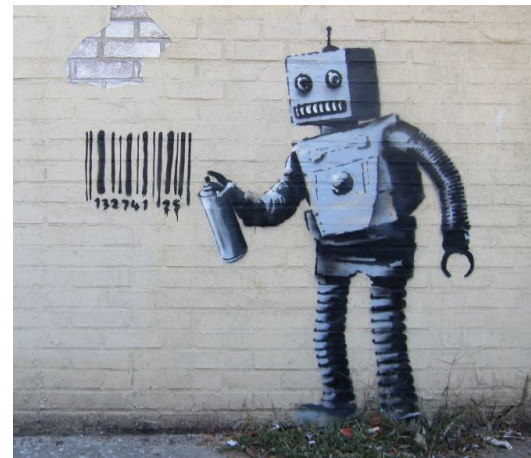


**Staatsbibliothek
zu Berlin**
Preußischer Kulturbesitz



Maschinelles Lernen in der Stabi Berlin

- Drittmittelprojekte
 - 2016—2024: **OCR-D** (DFG)
 - 2018—2021: **Qurator** (BMBF)
 - 2022—2025: **Mensch.Maschine.Kultur** (BKM)
- Team
 - 6x 100% Softwareentwicklung
 - 2x 50% Bibliotheks- und Informationswissenschaft
 - 1x 100% Forschungsdatenmanagement
- Hardware
 - 2x GPU Server NVIDIA Tesla A100
- Ziele
 - Forschung und Entwicklung zu Open Source Technologien für Digitalisierung von Kulturdaten
 - Erschließung und Durchsuchbarkeit für alle Inhalte, sowohl Text als auch Bild
 - Bereitstellung von Beständen als offene und maschinenlesbare Daten (**Collections as Data**)
 - Verantwortungsvoller Einsatz von KI und Kuratierung von digitalen Daten unter Berücksichtigung von problematischen Inhalten aus rechtlichen, sozialen und ethischen Perspektiven



CC-BY-SA Scott Lynch via [Flickr](#)

Mensch.Maschine.Kultur

- Das von BKM geförderte Projekt „Mensch.Maschine.Kultur“ besteht aus vier Teilprojekten, die aufeinander abgestimmt unterschiedliche Zielsetzungen verfolgen und mit den dafür geeigneten Verfahren kombinieren:
 - Teilprojekt 1 **“Intelligente Verfahren für die generische Dokumentenanalyse”** stellt Verfahren für die Dokumentenanalyse bereit, so dass qualitativ hochwertige Volltexte und Strukturdaten aus den verschiedenen digitalisierten Beständen gewonnen werden können (Text, Bild, Layout).
 - Teilprojekt 2 **“Bildanalyseinstrumente zur Erschließung des digitalen Kulturellen Erbes”** vertieft Arbeiten zur Bildanalyse durch Erkennung, Extraktion und Klassifizierung von digitalen Bildinhalten.
 - Teilprojekt 3 **“KI-unterstützte Inhaltsanalyse und Sacherschließung”** unterstützt die Expert:innen in den Fachabteilungen der Staatsbibliothek durch semi-automatisierte Verfahren bei der inhaltlichen Erschließung und bezieht dazu systematisch deren Expertise ein.
 - Teilprojekt 4 **“Datenbereitstellung und Kuratierung für KI”** bündelt und dokumentiert das digitalisierte kulturelle Erbe als Datensets spezifisch für den Einsatz von KI und erarbeitet Richtlinien darüber, wie problematische Inhalte erkannt und behandelt werden sollten.
- Alle Projektergebnisse (Software, Daten, Publikationen) werden offen zur Verfügung gestellt
- Für Ergebnisse und Updates zum Projektverlauf besuchen sie unsere Projektwebseite <https://mmk.sbb.berlin/> und das Blog <https://mmk.sbb.berlin/aktuelles/> bzw. für Software GitHub <https://github.com/qurator-spk> und Hugging Face <https://huggingface.co/SBB>

Analyse von Daten und Metadaten

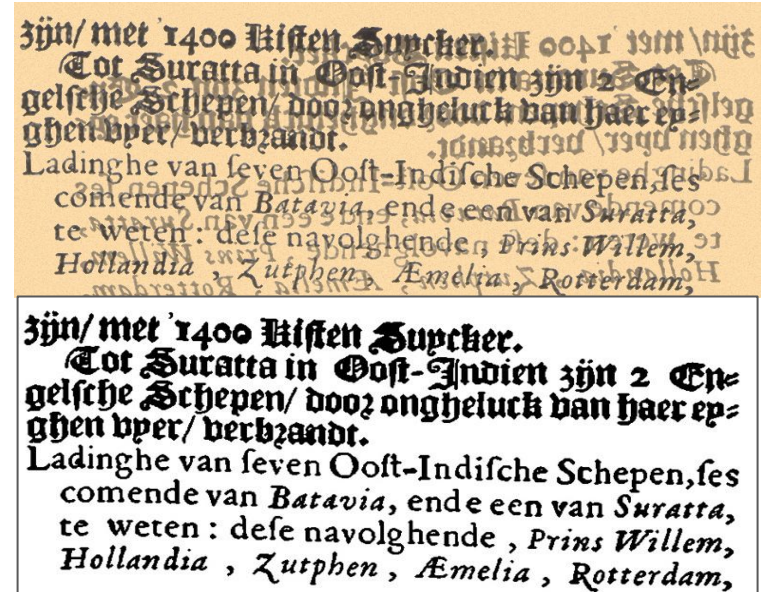
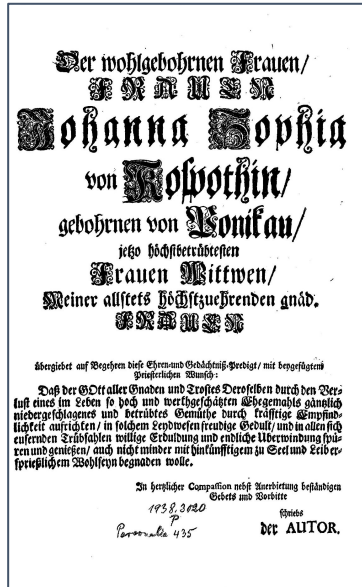
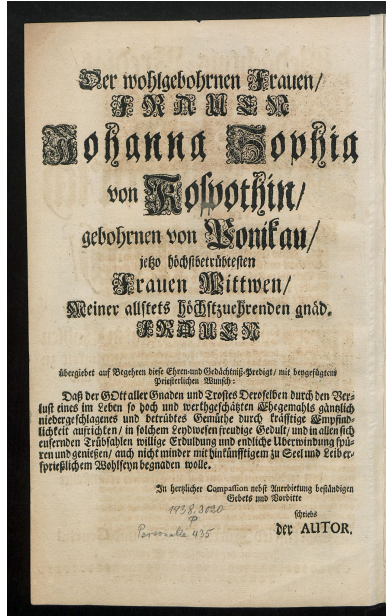
- Extraktion und Kombination diverser Datenquellen und deren Metadaten (METS, MODS, ALTO) für statistische Analyse mit Python Pandas
<https://github.com/qurator-spk/mods4pandas>

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 ...
...

- In der Praxis besteht das maschinelle Lernen oft zu 80% aus der Datenvorbereitung - Auswahl, Bereinigung, Umwandlung von Daten usw. - und nur zu 20% aus Entwicklung und Implementierung der Verfahren

Bildoptimierung und Binarisierung

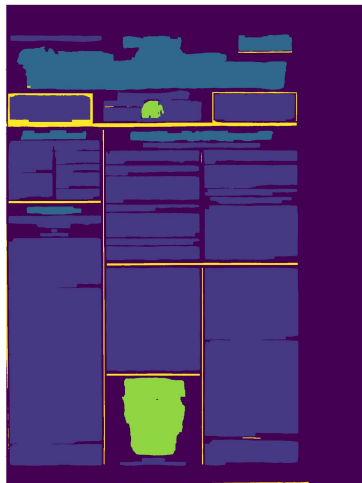
- *A hybrid CNN-Transformer model for Historical Document Image Binarization*, 2023.
<https://doi.org/10.1145/3604951.3605508> | https://github.com/qurator-spk/sbb_binarization



Layoutanalyse (Segmentierung)

- *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.*

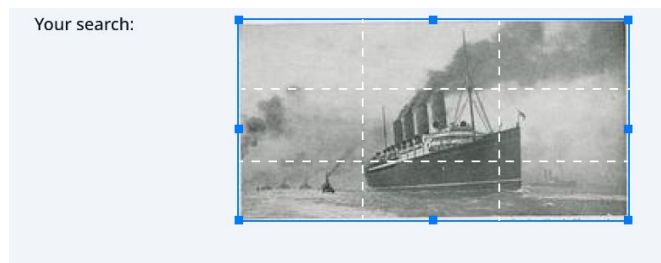
<https://doi.org/10.1145/3604951.3605516> | https://github.com/qurator-spk/sbb_images

Description Search Results

Image Description PPN

Big ship entering port

Enter image description. **English only!**



Images matching the description you entered:


Search Similar Images

View in Digitized Collections




Search Similar Images

View in Digitized Collections



Search Similar Images

View in Digitized Collections




Search Similar Images

View in Digitized Collections




Search Similar Images

View in Digitized Collections



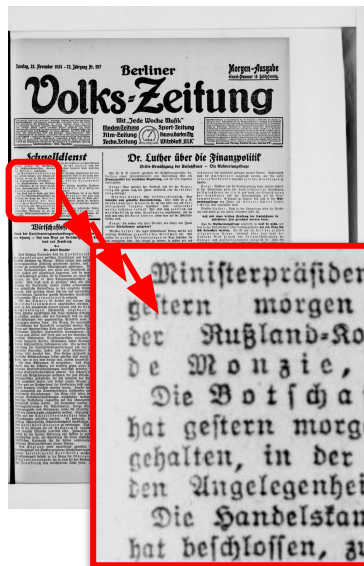
Search Similar Images

View in Digitized Collections



Texterkennung (OCR, HTR)

- *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 Herriot hat
Die Feierlichkeiten zur Ueberführung
geiern morgen den Vorsitzenden
der Autlanü-Kommission, Senator
d l M o l l z i e, empfangen.
Die Botschafterkonferenz
bat gestern morgen eine Sitzung ab,
gehalten, in der sie sich mit laufen-
ten Angelegenheiten beschäftigte.
Die Handelskammer von Bordeaux
bat beschloffen, zu der neuen franzö-
sischen Inlandsanleihe 1 Million
Francs zu zeichnen.
Aus Genf sind in Sofia zwei Dele-
gierte der Balkerbundskommission 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 Dölkerbundskommission ein-
getroffen.

Ministerpräsident Herriot hat
gestern morgen den Voritzenden
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 beschloffen, zu der neuen franzö-
sischen Inlandsanleihe 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-
fändige Völkerbundskommission ein-
getroffen.

OCR Evaluation und Qualitätssicherung

- Evaluation von OCR-Daten nach wissenschaftlichen Kriterien (CER, WER, Reading Order) basierend auf Ground Truth Daten

<https://github.com/qurator-spk/dinglehopper>

Character differences

20

rath mit einer P^ona ficali angehen worden, und folche durch des Hm. Graffen von Königsfeld Vor-
spruch, nur aus Gnaden nachgelaffen erhalten. Sondern man hat auch diefen 4. Wochen lang alle Abend bey der Inquifitt gantz allein gelaffen.

Binnen welcher ganzer Zeit der Schreiber Bredekaw befändig bey Ihme gewefen, und fich in

der am 1 3 ten Octobr. a. c. in Iudicio gegen feinen gewefenen Hm. introducirer Appellation defen Bey-
raths bedienet hat;

§. 33) Dabeneben lifft der Schreiber binnen diefer ganzen Zeit auf freyem Fuß geblieben, und

hat nicht nur durch feinen Confulenten, fondern auch, weilen der Inquifitt felbten in Ihrem Gefängnüß

fo viele Freyheit gelaffen worden, daß fie fremden Befuch von Ihren Anverwandten ohngehindert empfangen können, durch andere Perfonen fich mit ihr über alles, was Er oder fie dereinthen zu fagen hat-

20

rath mit einer P^ona ficali angehen worden, und folche durch des Hm. Graffen von Königsfeld Vor-
spruch, nur aus Gnaden nachgelaffen erhalten. Sondern man hat auch diefen 4. Wochen lang alle Abend bey der Inquifitt gantz allein gelaffen.

Binnen welcher ganzer Zeit der Schreiber Bredekaw befändig bey Ihme gewefen, und fich in

der am 1 3 ten Octobr. a. c. in Iudicio gegen feinen gewefenen Hm. introducirer Appellation defen Bey-
raths bedienet hat;

z3) Dabeneben lifft der Schreiber binnen diefer ganzen Zeit auf freyem Fuß geblieben, und

hat nicht nur durch feinen Confulenten, fondern auch, weilen der Inquifitt felbten in Jhrem Gefängnüß

fo viele Freyheit gelaffen worden, daß fie fremden Befuch von Jhren Anverwandten ohngehindert empfangen können, durch andere Perfonen fich mit ihr über alles, was Er oder fie dereinthen zu fagen hat-

Clemens Neudecker, Karolina Zaczyńska, 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 Nutzbarkeit. 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 in 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. zwischen Förderinstrumente führt dies jedoch zu einem hohen Grad an Heterogenität der Digitalisierungs- bzw. Textqualität sowie des Umfangs der strukturellen und semantischen Anreicherungen. Diese Heterogenität erschwert die Nachnutzung durch die Forschung, die neben einheitlichen

© Open Access. © 2023 Clemens Neudecker, Karolina Zaczyńska, Konstantin Baierer, u.a., publiziert von De Gruyter. <https://doi.org/10.1515/9783110693957-009>

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 Clauuser
Apostolos Antonacopoulos
Stefan Pletschacher
www.primaresearch.org
Pattern Recognition and Image Analysis Lab (PRIMAL)
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 when 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 analyse 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 Clauuser, Apostolos Antonacopoulos, and Stefan Pletschacher. 2023. A survey of OCR evaluation tools and metrics. In *The 4th International Workshop on Historical Document Imaging and Processing (HDIP '23)*, September 3–6, 2023, Louanane, Switzerland. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/364047.364048>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyright for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and a fee. Request permission from permissions@acm.org.

© 2023, Association for Computing Machinery.
ACM ISBN 978-1-60959-395-7, \$15.00.
<https://doi.org/10.1145/364047.364048>

1 INTRODUCTION

The efficient, transparent and informative evaluation of the quality of the results of Optical Character Recognition (OCR) is challenging in multiple respects. Established methods require Ground Truth (GT) data to serve as a reference for the desired result quality. Against the background of mass digitization¹, where millions of pages of documents are digitized and OCRed, this is neither feasible nor affordable. Especially in the context of historical documents, the creation of GT requires specialised skills and is a far too time-consuming to perform on a sufficiently large scale.

A further difficulty lies in the fact that standards or established conventions that provide clear and uniform guidelines for the creation of GT for historical documents are only partially available. There remain various un- or under-specified cases that can occur when assessing OCR quality. Examples include ligatures that can be recognised either as individual codepoints or as a combination of codepoints, characters that cannot be represented by a single codepoint, the encoding of special characters² that are not included in the Unicode standard and for which extensions such as MEI³ or other treatments from the Private Use Area⁴ must be used, and the treatment of punctuation and spaces. The OCR-D Ground Truth Guidelines [5] are an attempt to mediate between the OCR community and the needs of (scholarly) users of OCR results and to establish clearer specifications and guidelines.

In summary, established procedures and metrics for GT-based quality assessment of OCR results do not provide satisfactory answers when it comes to some of the more detailed questions that arise for historical documents. In addition, the extensive GT-based evaluation of large collections or an OCR in the context of mass digitization is not feasible. The question to which extent OCR confidence values and sample-based statistical evaluations can provide meaningful, reliable and comparable statements needs to be more systematically investigated. Finally, the quality of layout analysis seems to be insufficiently covered by established metrics.

This paper aims to raise and discuss issues of transparency and better direct comparability of OCR evaluations by identifying gaps and ambiguities in current methods and by putting the meaningfulness of OCR evaluation results more into the context of actual use cases for OCR results. The observations and analysis are drawn from

¹Google estimated in 2010 that there are around 190k unique books published at <http://books.google.com> in 2010 (books of world-wide origin and in countless digital and digital editions in October 2019). <https://www.blog.google/products/search/3-years-google-books/>

²https://en.wikipedia.org/wiki/Unicode_extensions, <https://www.unicode.org/review/>

³Unicode Standard, Free Indivisors. <https://ikl.ac.uk/unicode/>

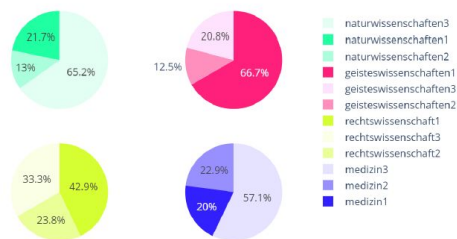
⁴Unicode Standard, Chapter 16: Special Area and Format Characters.

Semi-automatische Sacherschließung

- Training von **Annif** Modellen auf Grundlage von ARK und Basisklassifikation
- Bereitstellung von **Trainingsdaten** und trainierten **Modellen**
- Quantitative und qualitative Evaluation der maschinellen Sacherschließung

	Projekt ID	Anzahl Texte/ Titel	P*	R*	F1*	NDCG*	*eval Parameter	F1@5	NDCG@5
Volltexte	ark-mlm-de-18	ca. 1100	0.0620	0.0775	0.0206	0.0523		0.0257	0.0473
	ark-tfidf-de-18	ca. 1100	0.2394	0.1933	0.2081	0.2036	11/t .45	0.1535	0.3152
	ark-omikujj-de-18	ca. 1100	0.3404	0.3991	0.3452	0.3938	12/t .05	0.2216	0.4866
	ark-omikujj-de	11753	0.2425	0.4538	0.3126	0.4415	12/t 0	0.1846	0.4755
	ark-fasttext-de-18	ca. 1100	0.2958	0.2285	0.2496	0.2434	11/t 0	0.1497	0.3275
Titel-/Metadaten	ark-omikujj-de-title-only	787.721	0.4853	0.4582	0.4669	0.4678	11/t 0	0.3771	0.5243
	ark-omikujj-de-title-content	787.721	0.4861	0.4587	0.4675	0.4683	11/t 0	0.2258	0.5703
	ark-omikujj-lat-title-content	146.522	0.5755	0.4916	0.5176	0.5153	11/t 0	0.2676	0.6305
	ark-omikujj-multilingual-title-content	2.518.792	0.4779	0.4521	0.4604	0.4639	11/t 0	0.2197	0.5610
	ark-omikujj-fre-title-content	87.195	0.4228	0.4044	0.4103	0.4169	11/t 0	0.1885	0.4939
	ark-omikujj-eng-title-content	85.124	0.3838	0.3694	0.3741	0.3795	11/t 0	0.1839	0.4702

Qualitative Bewertung von je 5 Vorschlägen für insgesamt 103 Titel durch Kolleg:innen aus der Sacherschließung
Einteilung in die Kategorien: (1) sehr nützlich, (2) nützlich und (3) nicht nützlich



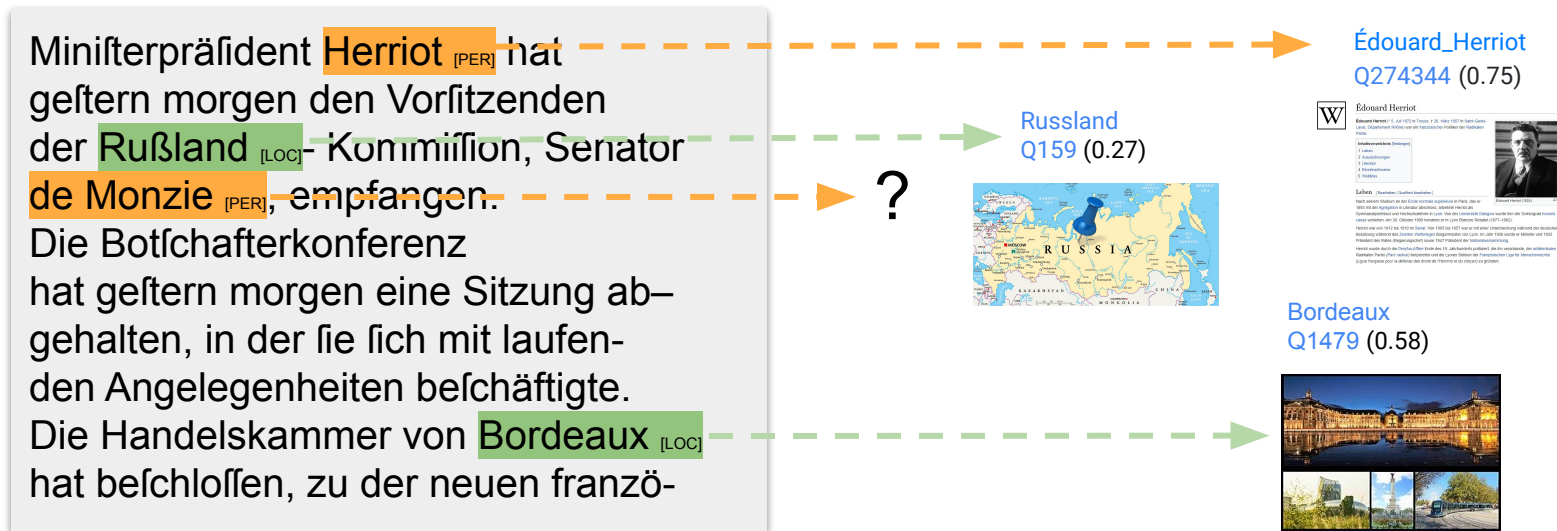
- **Verteilung der Evaluationsergebnisse**
nach der jeweils besten vergebenen Kategorie pro Titel und nach Fachbereichen
- Naturwissenschaften, n=23, Modell: [de](#)
 - Geisteswissenschaften, n=24, Modell: [multilingual](#)
 - Rechtswissenschaft, n=21, Modell: [de](#)
 - Medizin, n=35, Modell: [de](#)

Named Entity Recognition und 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



Datensets und Richtlinien für Datendokumentation

- *Datasheets for Digital Cultural Heritage Datasets, 2023.*

<https://doi.org/10.5334/johd.124> | <https://doi.org/10.5281/zenodo.8375034>

The screenshot shows the 'Europeana pro' logo and the title 'Datasheets for digital cultural heritage Working Group'. A badge indicates the page was updated on Monday, November 6, 2023. Below the title, a paragraph states: 'This Working Group, set up within the Europeana Research Community and EuropeanaTech Community, works to adapt the concept of datasheets for the cultural heritage sector.' The page features a navigation bar with 'Hugging Face' and various menu items like 'Models', 'Datasets', 'Spaces', 'Docs', 'Solutions', 'Pricing', 'Log In', and 'Sign Up'. The main content area displays the 'Staatsbibliothek zu Berlin - Preußischer Kulturbesitz' profile, including its logo, website URL, and a 'Request to join this org' button. Below this, there are sections for 'AI & ML interests' and 'Team members'. A prominent 'Organization Card' for the Staatsbibliothek zu Berlin is shown, describing it as one of the largest scientific universal libraries in Germany, digitizing its collections and making them available online. It mentions research projects like 'Quator' and 'Mensch.Maschine.Kultur'.

Henk Alkemade, Steven Claeysens, Giovanni Colavizza, Nuno Freire, Alba Irollo, Jörg Lehmann, Clemens Neudecker, Giulia Osti, Daniel van Strien

Template: Datasheet for Digital Cultural Heritage Datasets

Version 1 – September 2023

The superscripts added to section headings refer to items in the bibliography at the very end, where every section is thoroughly discussed, explained and further questions can be found. The main structure follows Gebru's (2021) and Pushkarna's templates (2022).

Motivation^{b,f}

[Clearly articulate the reasons for creating the dataset and promote transparency about funding interests. Also provide a brief descriptive overview of the dataset ('at a glance'). For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organisation)? Who funded the creation of the dataset?]

Dataset Description^c

Homepage [Add homepage URL here if available.]

Repository [E.g., if the dataset is hosted on GitHub or has a GitHub homepage, add URL here.]

Paper [If the dataset was introduced by a paper or there was a paper written describing the dataset, add URL here.]

Point of Contact [If known, name and email of at least one person the reader can contact for questions about the dataset.]

Dataset Summary^e

[Briefly summarise the dataset, its intended use and the supported tasks. Give an overview of how and why the dataset was created. The summary should explicitly mention the languages present in the dataset (possibly in broad terms, e.g. translations between several pairs of European languages), and describe the domain, topic, genre covered, keywords, and other relevant metadata.]

III. Herausforderungen und Chancen

Herausforderungen von Kulturdaten bei Gebrauch von KI

- Digitalisierte Daten und Metadaten liegen nicht in der für das Trainieren von KI erforderlichen Form und den entsprechenden Formaten vor und müssen daher zunächst evaluiert, konvertiert und/oder (von Expert:innen) transkribiert und annotiert werden
- Aufgrund des geltenden Urheberrechts werden vor allem historische Werke digitalisiert, diese enthalten historische Begriffe und Schreibweisen, für die KI erst angepasst oder trainiert werden muss (was für die KI-Industrie wirtschaftlich nicht relevant ist)
- Kulturelle Kontexte (z.B. Zeit, Ort) müssen bei kulturellem Erbe und Anwendung von KI berücksichtigt werden - wie lautet z.B. die Antwort auf die Frage "Berlin ist die Hauptstadt von ???" während 1618-1701, 1701-1918, 1918-33, 1933-45, 1945-49, 1949-89, 1990-heute?
- Ethisch, sozial oder rechtlich problematische Inhalte (wie Kolonialismus, Nationalsozialismus oder die Unterrepräsentation von marginalisierten Gruppen) in den Quelldaten müssen identifiziert und mit entsprechender Vorsicht und Sorgfalt behandelt und kontextualisiert werden
- Es wird spezielle Hardware (GPUs) und Fachwissen (Einstellung von Personal mit den entsprechenden Fähigkeiten) benötigt - der öffentliche Sektor ist hier weit von den Ressourcen entfernt, die in privatwirtschaftlichen Unternehmen zur Verfügung stehen

Chancen für Kulturerbeeinrichtungen im Umgang mit KI

- Die Anwendung von KI biete eine Vielzahl von Möglichkeiten für effizientere Erschließung, Analyse und Anreicherung von digitalisierten Kulturdaten und eröffnet Möglichkeiten für neue Dienste für Nutzende und für die wissenschaftliche Forschung
- Die Sammlungen wurden über Jahrhunderte hinweg zusammengetragen, von Fachleuten aufgrund ihrer Relevanz ausgewählt und nach gemeinsamen Standards kuratiert
- Aufgrund der fortschreitenden Massendigitalisierung verfügen Kulturerbeeinrichtungen über große und ständig wachsende Mengen an (meist offenen) und qualitativ hochwertigen Daten, die für das Training und die Verbesserung von KI-Modellen zur Verfügung stehen
- Die Expert:innen in den Fachabteilungen der Kulturerbeeinrichtungen verfügen über ein enormes tiefes Wissen über die Sammlungen und deren Inhalte, von dem KI lernen und profitieren kann
- Im Allgemeinen ist das Qualitätsbewusstsein und die Sensibilität bei der Erstellung, Pflege und Nutzung von Daten stark ausgeprägt (etwa im Vergleich zu großen Technologieunternehmen)
- Als öffentliche Einrichtungen und Dienstleister für die Forschung werden Transparenz, Datenschutzrecht und Verantwortung im Umgang mit Daten ernst genommen - auch über sehr lange Zeiträume hinweg und ohne die Notwendigkeit, Fähigkeiten zu übertreiben und Probleme der KI zu verschleiern
- Erfahrungen können geteilt und Ressourcen gebündelt werden - wir sind keine Konkurrenten

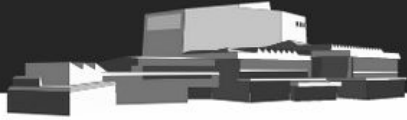
“In closing, we advocate for a turn in the culture towards **carefully collected** datasets, **rooted in their original contexts**, distributed only in ways that **respect the intellectual property and privacy rights** of data creators and data subjects, and constructed **in conversation with the relevant 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.”

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, Alex Hanna (2021). Data and its (dis)contents: A survey of dataset development and use in machine learning research. <https://doi.org/10.1016/j.patter.2021.100336>



Culture for AI

AI for Culture



Staatsbibliothek
zu Berlin
Preußischer Kulturbesitz



Danke für die Aufmerksamkeit! Fragen?

Diese Folien finden Sie auch online:

<http://sbb.berlin/o1npe>

