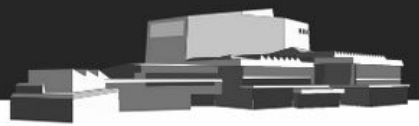


Staatsbibliothek
zu Berlin
Preußischer Kulturbesitz

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

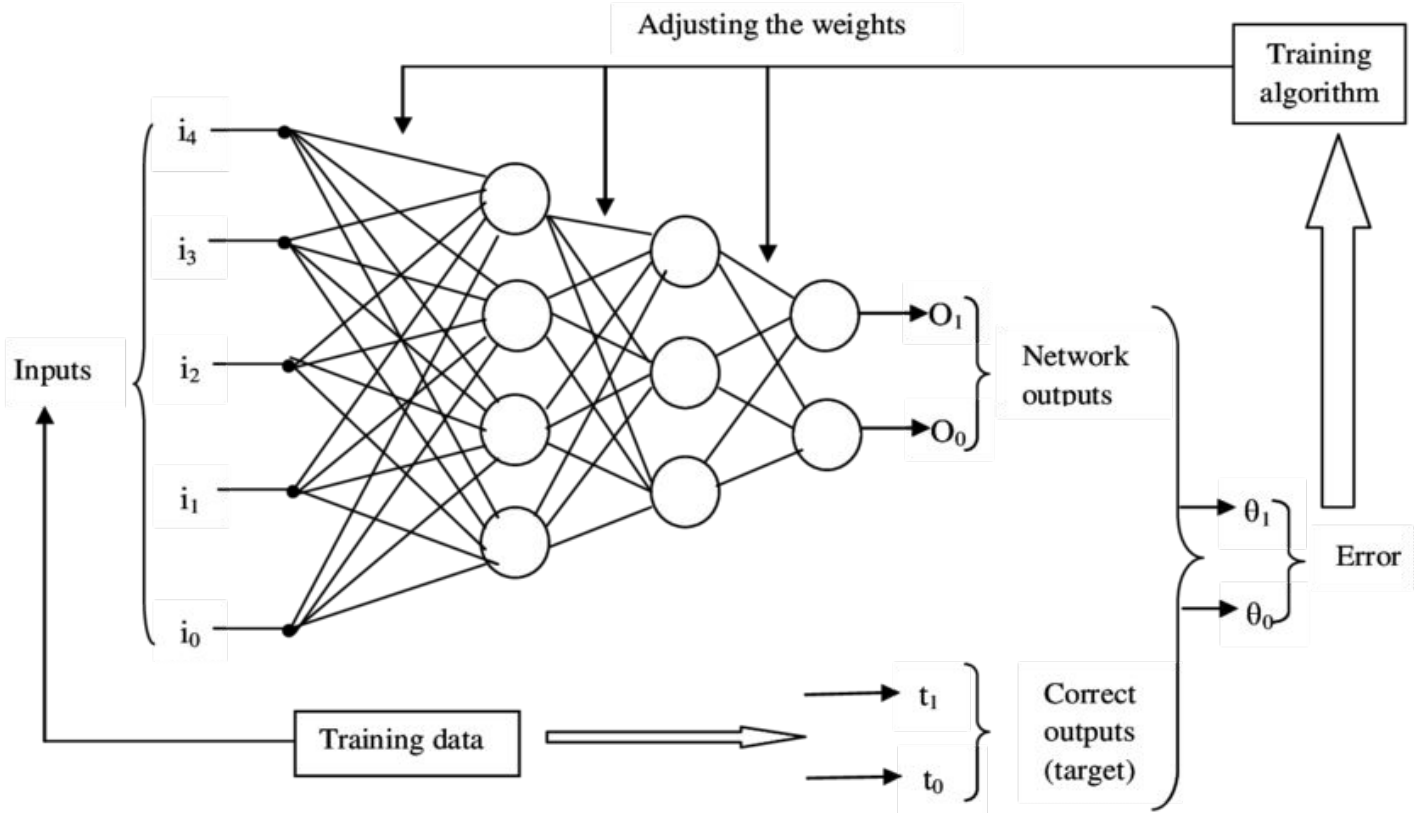
Was genau ist Künstliche Intelligenz (bzw. Deep Learning)?

- 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?

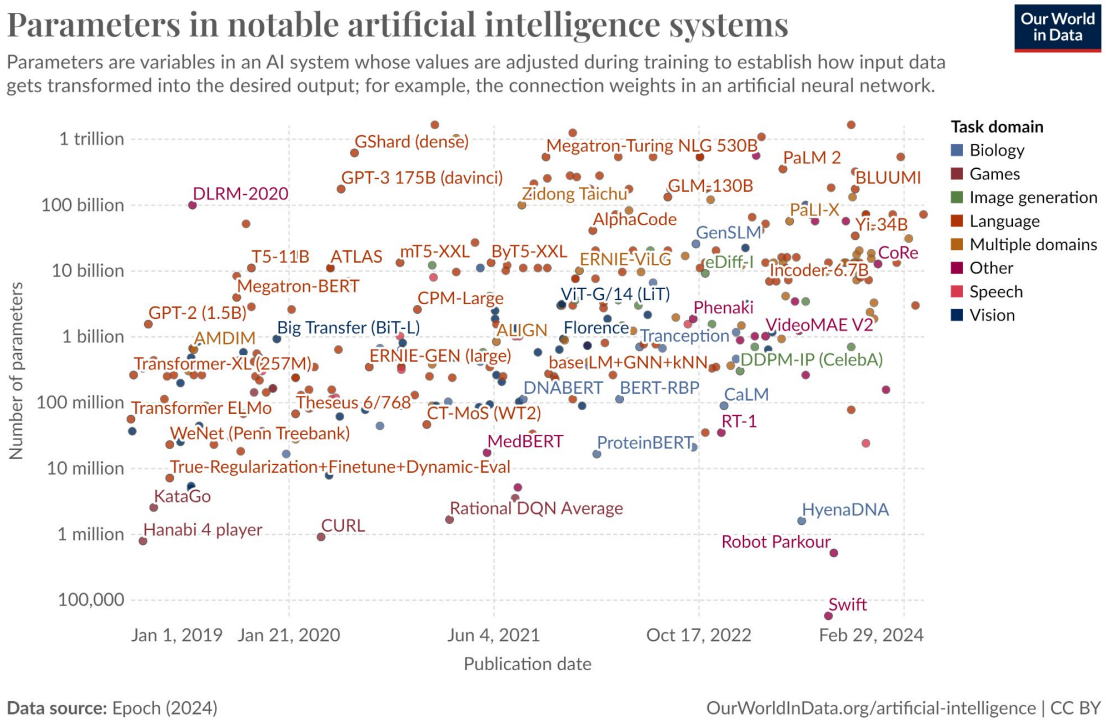
Beispiele

- 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



Consumption	CO_2e (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 IIIF 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

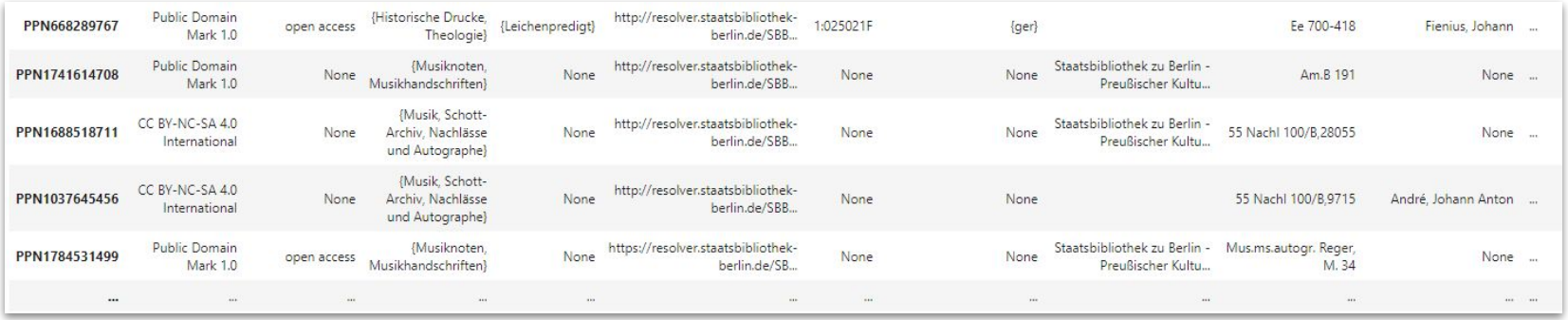
- Drittmittelprojekte
  - 2016—2024: **OCR-D** (DFG)
  - 2018—2021: **Qurator** (BMBF)
  - 2022—2025: **Mensch.Maschine.Kultur** (BKM)
- Team
  - 6x 100% Softwareentwicklung
  - 2x 50% Bibliotheks- und Informationswissenschaft
  - 1x 100% Forschungsdatenmanagement
- Hardware
  - 2x GPU Server NVIDIA 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

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

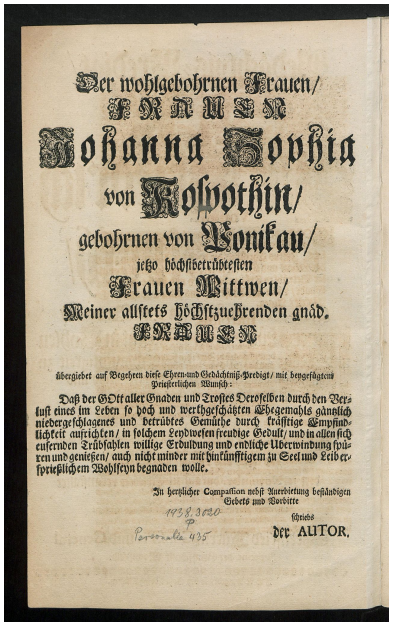


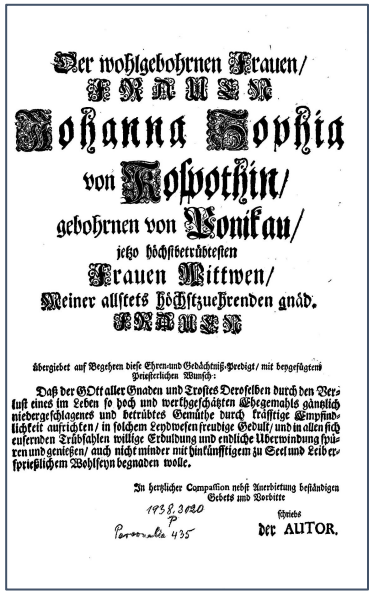
PPN668289767	Public Domain Mark 1.0	open access	{Historische Drucke, Theologie}	{Leichenpredigt}	http://resolver.staatsbibliothek-berlin.de/SBB...	1:025021F	{ger}		Ee 700-418	Fienius, Johann	...
PPN1741614708	Public Domain Mark 1.0	None	{Musiknoten, Musikhandschriften}	None	http://resolver.staatsbibliothek-berlin.de/SBB...	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	...
...	...	...	...	...	...	...	...	...	...	...	...

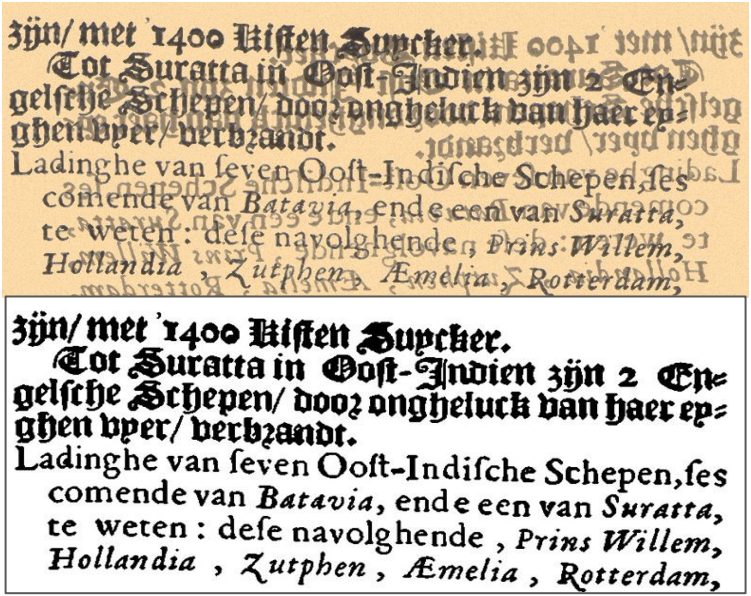
- In der Praxis besteht das maschinelle Lernen oft zu 80% aus der Datenvorbereitung - Auswahl, Bereinigung, Umwandlung von Daten usw. - und nur zu 20% aus Entwicklung und Implementierung der Verfahren

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**

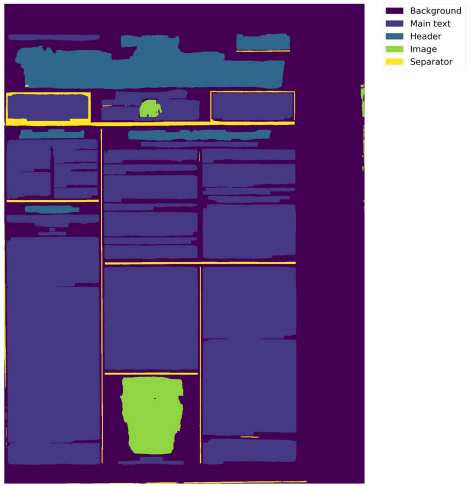






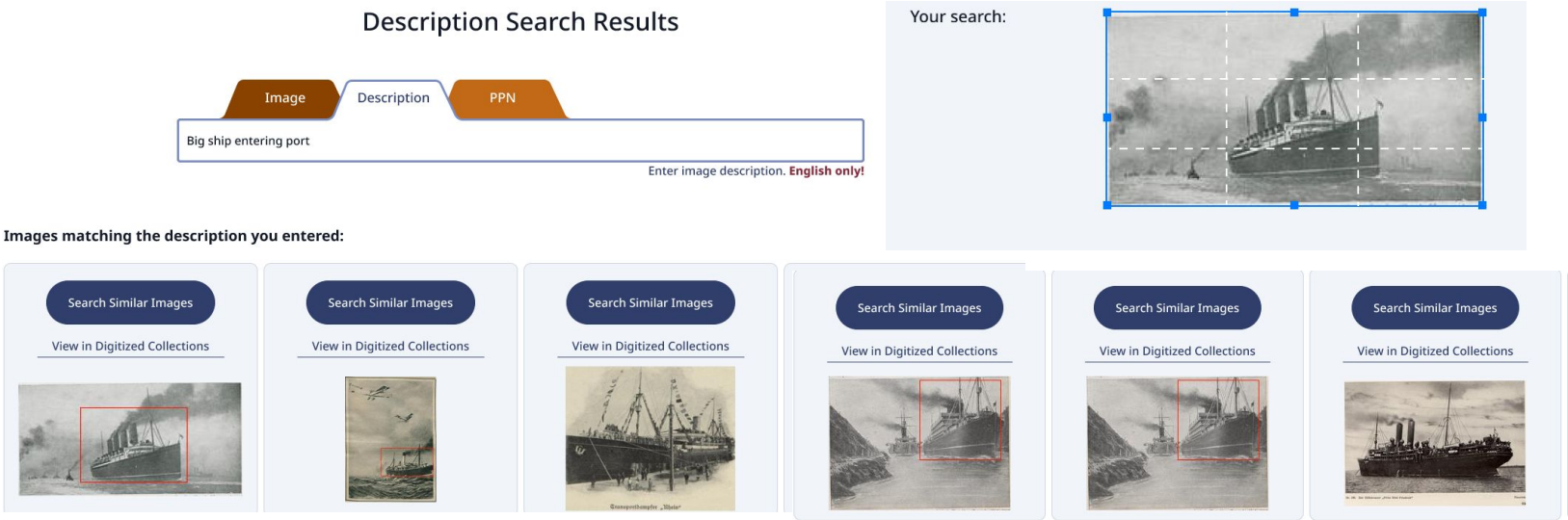
Layoutanalyse (Segmentierung)

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



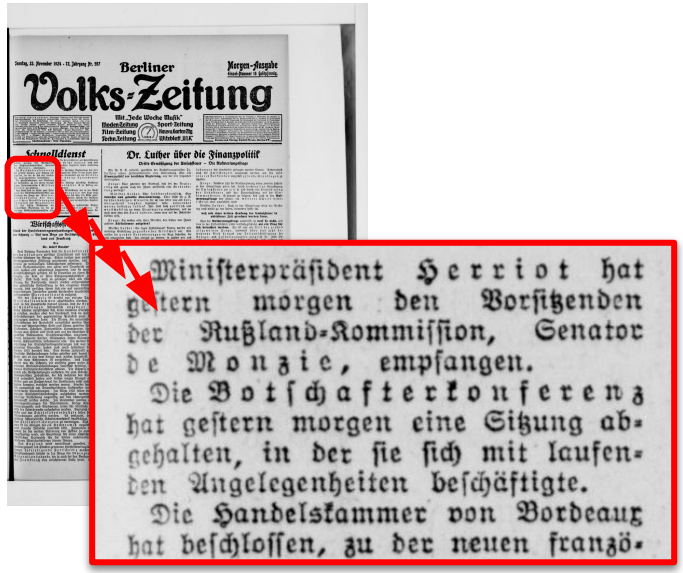
Bildähnlichkeitssuche und Text-Bild-Suche

- *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**



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**

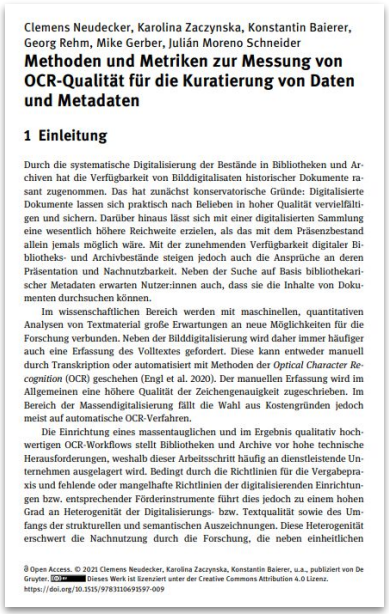


Ministerpräsident Hcrri 0 t hat
Die Feierlichkeiten zur UcbcrfUhrung
geilern morgen de» Vorsitzenden
der Autzlanü-Komniission, Cenator
dl M 0 ll z i e, empfangen.
Die Botschafterkonferenz
bat gestern morgen eine Citzung ab,
gehalten, in der sie sich mit laufen-
ten Angelegenheiten beschäftigte.
Die Handelskammer 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

- Evaluation von OCR-Daten nach wissenschaftlichen Kriterien (CER, WER, Reading Order) basierend auf Ground Truth Daten
  https://github.com/qurator-spk/dinglehopper

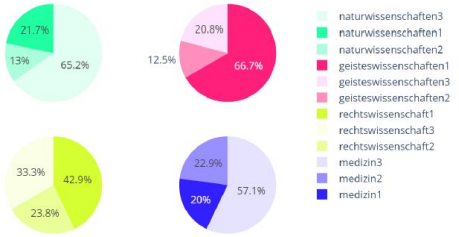


Semi-automatische Sacherschließung

- 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
Vollexte	ark-mllm-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	l 1/t .45	0.1535	0.3152
	ark-omikuji-de-18	ca. 1100	0.3404	0.3991	0.3452	0.3938	l 2/t .05	0.2216	0.4866
	ark-omikuji-de	11753	0.2425	0.4538	0.3126	0.4415	l 2/t 0	0.1846	0.4755
	ark-fasttext-de-18	ca. 1100	0.2958	0.2285	0.2496	0.2434	l 1/t 0	0.1497	0.3275
Titel-/Metadaten	ark-omikuji-de-title-only	787.721	0.4853	0.4582	0.4669	0.4678	l 1/t 0	0.3771	0.5243
	ark-omikuji-de-title-content	787.721	0.4861	0.4587	**0.4675**	0.4683	l 1/t 0	0.2258	0.5703
	ark-omikuji-lat-title-content	146.522	0.5755	0.4916	**0.5176**	0.5153	l 1/t 0	0.2676	0.6305
	ark-omikuji-multilingual-title-content	2.518.792	0.4779	0.4521	**0.4604**	0.4639	l 1/t 0	0.2197	0.5610
	ark-omikuji-fre-title-content	87.195	0.4228	0.4044	**0.4103**	0.4169	l 1/t 0	0.1885	0.4939
	ark-omikuji-eng-title-content	85.124	0.3838	0.3694	**0.3741**	0.3795	l 1/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

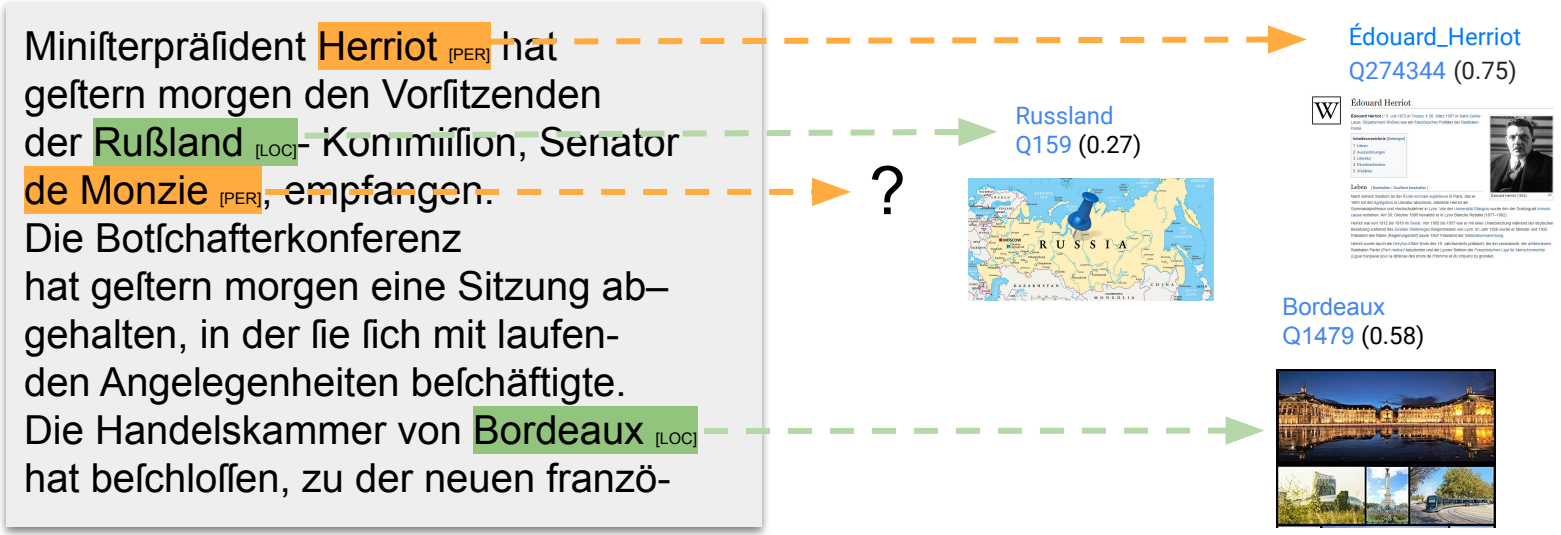


Verteilung der Evaluationsergebnisse
nach der jeweils besten vergebenen Kategorie pro Titel und nach Fachbereichen

- Naturwissenschaften, n=23, Modell: de
- Geisteswissenschaften, n=24, Modell: multilingual
- Rechtswissenschaft, n=21, Modell: de
- Medizin, n=35, Modell: de

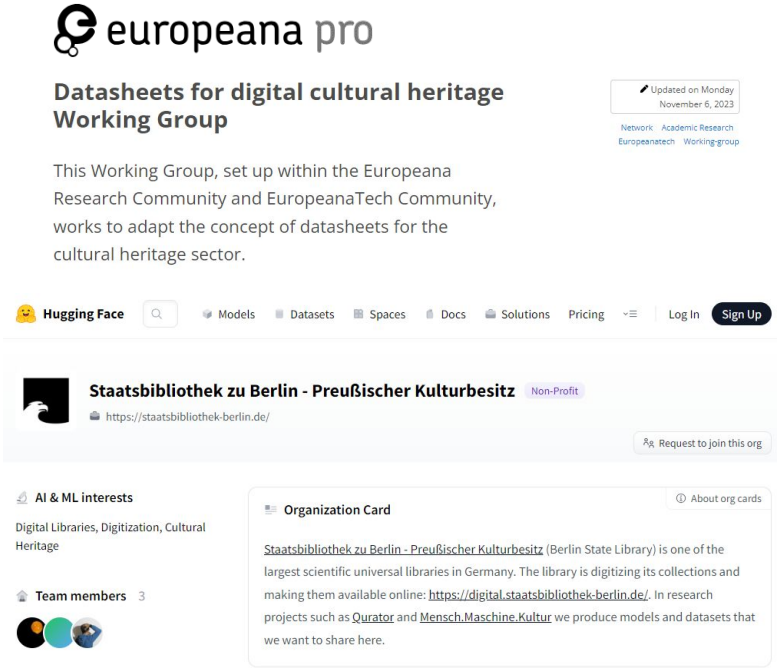
Named Entity Recognition und Entity Linking

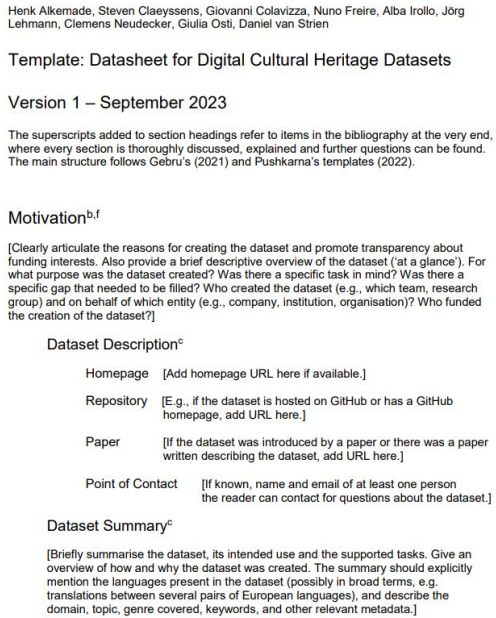
- *BERT for Named Entity Recognition in Contemporary and Historic German*, 2019.
  https://konvens.org/proceedings/2019/papers/KONVENS2019_paper_4.pdf |
  https://github.com/qurator-spk/sbb_ner | **https://github.com/qurator-spk/sbb_ned**



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**





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

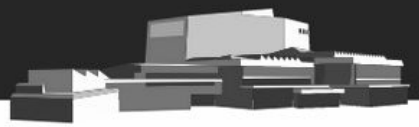
Chancen für Kulturerbeeinrichtungen im Umgang mit KI

- 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**

Many thanks to Steven Claeyssens for this excellent meme!



Staatsbibliothek
zu Berlin
Preußischer Kulturbesitz

Danke für die Aufmerksamkeit!
Fragen?

Diese Folien finden Sie auch online:

http://sbb.berlin/o1npe