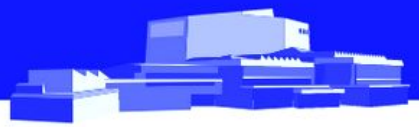


Staatsbibliothek
zu Berlin
Preußischer Kulturbesitz

Einsatzszenarien von KI für die Texterkennung und Entitätenerkennung an der Staatsbibliothek zu Berlin

Clemens Neudecker | Staatsbibliothek zu Berlin — Preußischer Kulturbesitz

5. KOBV-Fachkolloquium 2022 | 28. November 2022 | Zuse Institut Berlin

KI-Projekte @ SBB

- QURATOR (BMBF)
  - https://qurator.ai/
  - 11/2018 — 10/2021
  - 3x 100% E13 (Machine Learning Engineer)

- Mensch.Maschine.Kultur (BKM)
  - 07/2022 — 06/2025
  - 4x 100% E13 (Machine Learning Engineer)
  - 1x 100% E13 (Forschungsdatenmanagement)
  - 2x 50% E13 (Bibliothekar*in)
  - 1x 50% E11 (Projektmanagement)

QURATOR

- 10 Projektpartner mit eigenständigen Teilprojekten
- 6 KMUs, Fraunhofer Fokus, DFKI, Wikimedia, SPK (ausführend: SBB)
- Teilprojekt 10: “Kuratierungstechnologien für das digitale kulturelle Erbe”
- Regionaler Wachstumskern für den Standort Berlin-Brandenburg
- Ergebnisse der SBB
  - Open Source Software https://github.com/qurator-spk
  - KI Modelle https://qurator-data.de/ bzw. https://huggingface.co/SBB (im Aufbau)
  - Datensets https://zenodo.org/communities/stabi
  - Wiss. Publikationen https://github.com/qurator-spk/publications
  - Demonstratoren https://ravius.sbb.berlin/ (im Aufbau)

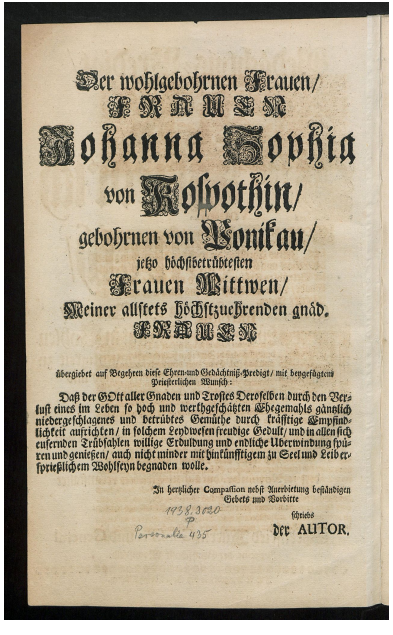
QURATOR: mods4pandas

- Werkzeuge für Metadatenanalyse, Merkmalsextraktion und statistische Auswertungen von Digitalisaten
  - https://github.com/qurator-spk/mods4pandas
  - Datenbasis
    - METS, MODS, ALTO
    - Aus Images extrahierte Merkmale (z.B. Anzahl Spalten, Bildeigenschaften)
  - Stellt Metadaten und extrahierte Merkmale als pandas Dataframe für ML zur Verfügung
  - Grundlage für z.B. Jupyter Notebooks, Datenvisualisierungen

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	...
...	...	...	...	...	...	...	...	...	...	...	...

QURATOR: Binarisierung

- Binarisierung für Texterkennung
  - https://github.com/qurator-spk/sbb_binarization
  - Hybrid CNN + Transformer



Der wohlgebohrnen Frauen/
FRAUEN
Johanna Sophia
von Kospoth/
gebohrnen von Bonkau/
jetzo höchstbetrübtesten
Frauen Wittwen/
Meiner allstets höchstzuehrenden gnäd.
FRAUEN

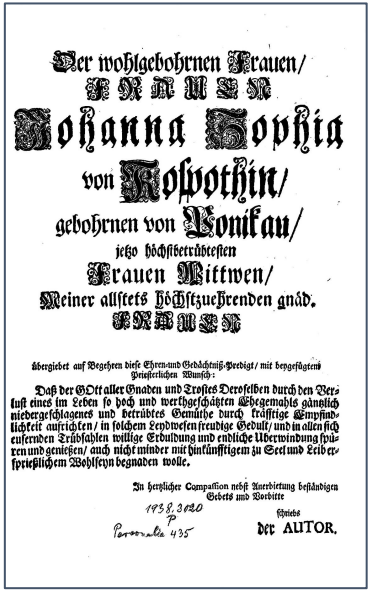
übergiebet auf Begehren diese Ehren- und Gedächtniß-Predigt/ mit beygefügtem Priesterlichen Wunsch:

Daß der Gott aller Gnaden und Trostes Derselben durch den Verlust eines im Leben so hoch und werthgeschätzten Ehegemahls gänzlich niedergeschlagenes und betrübtes Gemüthe durch kräfftige Empfindlichkeit aufrichten/ in solchem Leydwesen freudige Geduld/ und in allen sich eusernden Trübsahlen willige Erduldung und endliche Überwindung spüren und genießen/ auch nicht minder mit hinkünfftigem zu Seel und Leib ersprießlichem Wohlseyn begnaden wolle.

In herzlicher Compassion nebst Anerbietung beständigen Gebets und Wohlwitte

schrieb
der AUTOR.

1938.3020
P
Porzellan 435



Der wohlgebohrnen Frauen/
FRAUEN
Johanna Sophia
von Kospoth/
gebohrnen von Bonkau/
jetzo höchstbetrübtesten
Frauen Wittwen/
Meiner allstets höchstzuehrenden gnäd.
FRAUEN

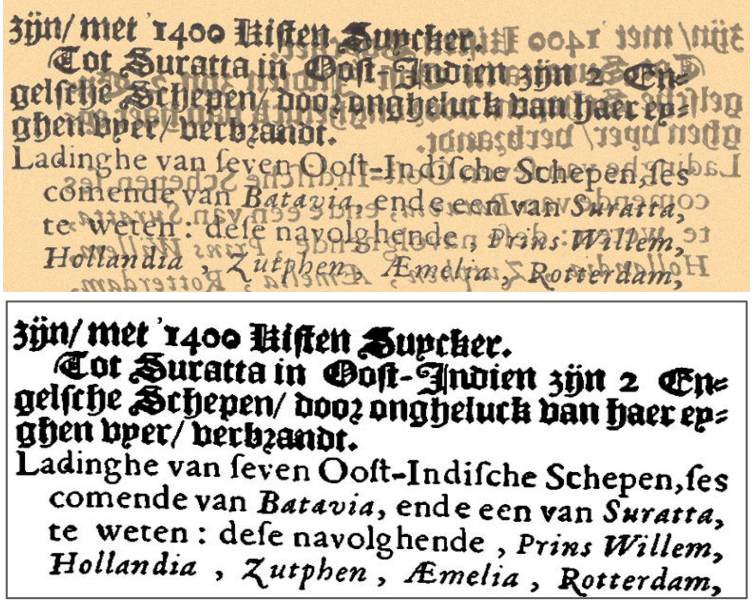
übergiebet auf Begehren diese Ehren- und Gedächtniß-Predigt/ mit beygefügtem Priesterlichen Wunsch:

Daß der Gott aller Gnaden und Trostes Derselben durch den Verlust eines im Leben so hoch und werthgeschätzten Ehegemahls gänzlich niedergeschlagenes und betrübtes Gemüthe durch kräfftige Empfindlichkeit aufrichten/ in solchem Leydwesen freudige Geduld/ und in allen sich eusernden Trübsahlen willige Erduldung und endliche Überwindung spüren und genießen/ auch nicht minder mit hinkünfftigem zu Seel und Leib ersprießlichem Wohlseyn begnaden wolle.

In herzlicher Compassion nebst Anerbietung beständigen Gebets und Wohlwitte

schrieb
der AUTOR.

1938.3020
P
Porzellan 435

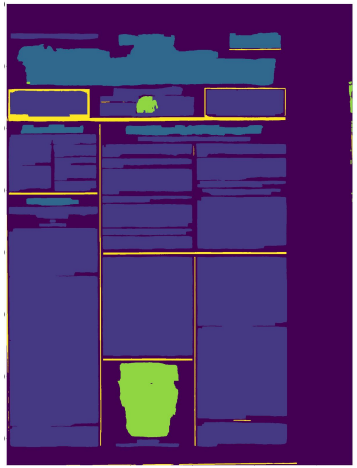


zijn/ met 1400 Kisten Suycker.
Tot Suratta in Oost-Indien zijn 2 Engelsche Schepen/ dooz ongheluck van haer eghen vyer/ verbzandt.
Ladinghe van seven Oost-Indische Schepen, ses comende van *Batavia*, ende een van *Suratta*, te weten: dese navolghende, *Prins Willem*, *Hollandia*, *Zutphen*, *Ämelia*, *Rotterdam*,

zijn/ met 1400 Kisten Suycker.
Tot Suratta in Oost-Indien zijn 2 Engelsche Schepen/ dooz ongheluck van haer eghen vyer/ verbzandt.
Ladinghe van seven Oost-Indische Schepen, ses comende van *Batavia*, ende een van *Suratta*, te weten: dese navolghende, *Prins Willem*, *Hollandia*, *Zutphen*, *Ämelia*, *Rotterdam*,

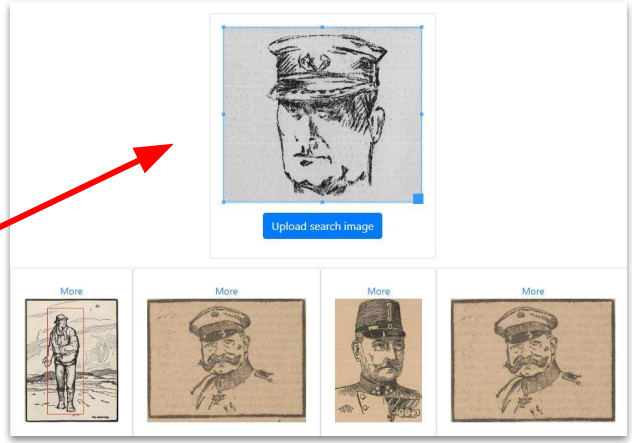
QURATOR: Eynollah

- Bildoptimierung und Layoutanalyse, Reading Order
  - https://github.com/qurator-spk/eynollah
  - Pixelweise Segmentierung ResNet-U-Net + Transformer + Heuristiken



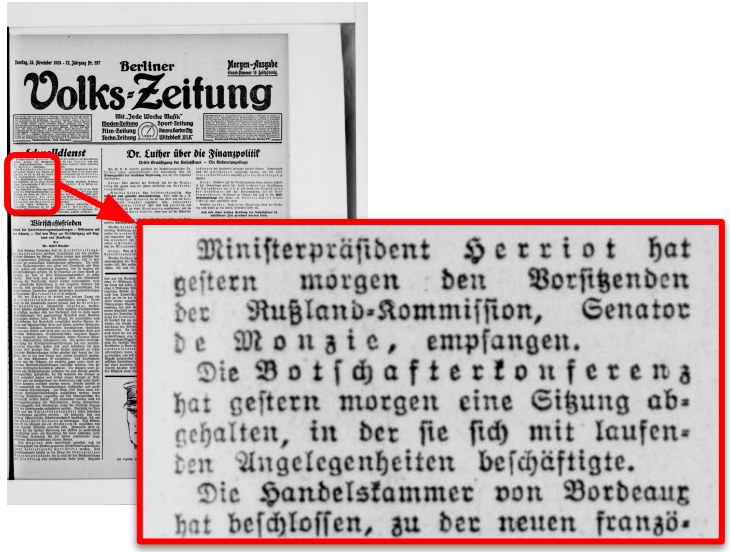

QURATOR: Bildähnlichkeitssuche

- Bildähnlichkeitssuche und Bildklassifikation
  - https://github.com/qurator-spk/sbb_images (Demo)
  - GoogleNet/Inception + YOLO + Saliency + Approximate Nearest Neighbour



QURATOR: Texterkennung (OCR)

- In Zusammenarbeit mit OCR-D wurde die Fehlerrate um bis zu 90% reduziert
  - https://github.com/qurator-spk/ocrd_calamari
  - CNN + LSTM + CTC

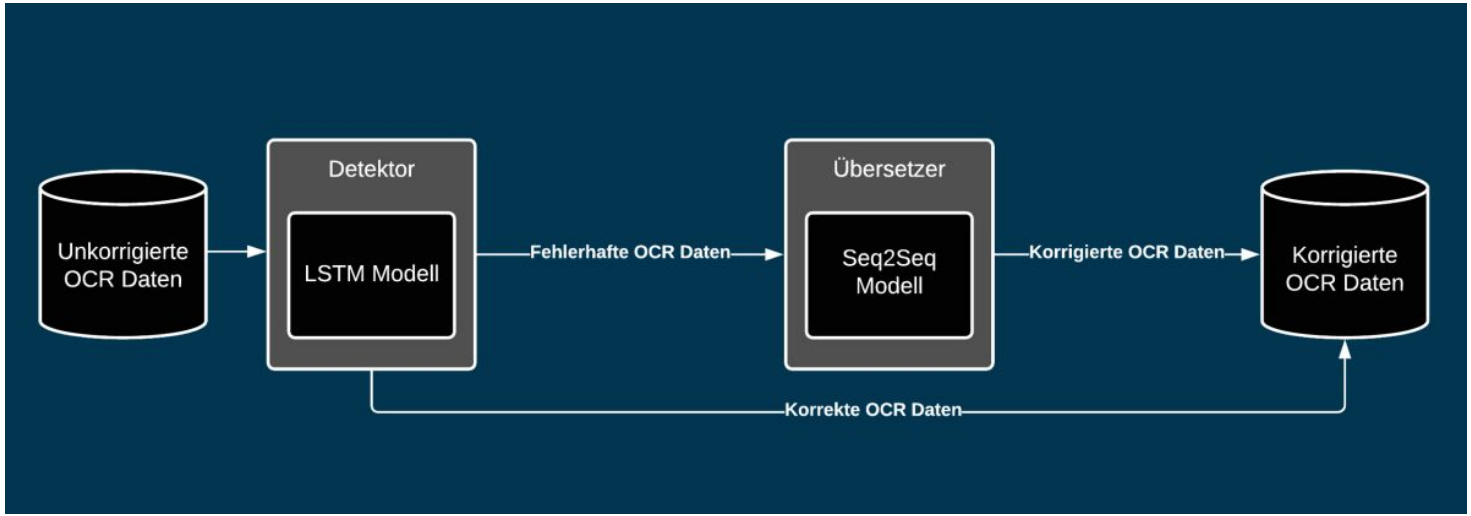


Ministerpräsident Hcrri 0 t hat
Die Feierlichkeiten zur UcbcrfUhrung
geilern morgen de» Vorsitzenden
der Autzlanü-Komniission, Cenator
dl M 0 II z i e, empfangen.
Die Botschafterkonferenz
bat gestern morgen eine Citzung ab,
gehalten, in der sie sich mit laufenten Angelegenheiten beschäftigte.
Die Handelskammer von Bordeaux
bat defchloffen, zu der neuen franzö_
sischen Inlandsanleihe 1 Million
Francs zu zeichne».
Aus Genf sind in Sofia zwei Delegierte 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 eingetroffen.

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 laufenden Angelegenheiten beſchäftigte.
Die Handelskammer von Bordeaux
hat beſchloſſen, zu der neuen franzöſiſchen Inlandsanleihe 1 Million
Francs zu zeichnen.
Aus Genf ſind in Sofia zwei Delegierte der Völkerbundskommiſſion zur
Prüfung der Frage der Maſſenauswanderung der bulgariſchen Bevölkerung aus
Thrazien und Mazedonien
und der letzten Beſchwerde der
bulgariſchen Regierung an die zu–
ſtändige Völkerbundskommiſſion eingetroffen.

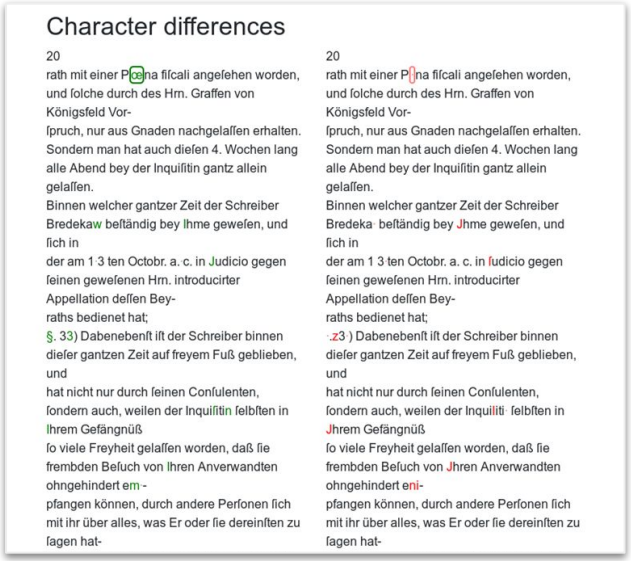
QURATOR: Nachkorrektur (OCR)

- Automatisierte Nachkorrektur von OCR Resultaten
  - https://github.com/qurator-spk/sbb_ocr_postcorrection (Paper)
  - Sequence-to-Sequence LSTM



QURATOR: Evaluierung (OCR)

- Evaluierung von OCR Qualität nach CER und WER
  - https://github.com/qurator-spk/dinglehopper
  - Alignierung + Levenshtein Distanz + Visual Diff




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

1 INTRODUCTION


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.

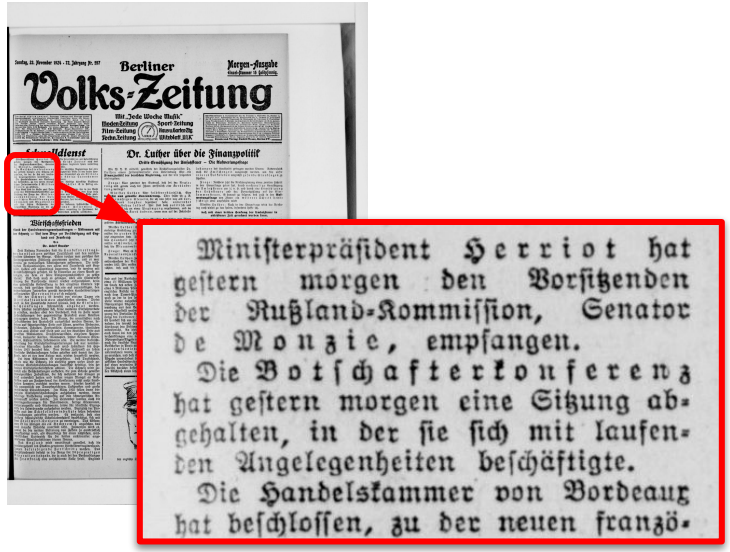
Im wissenschaftlichen Bereich werden mit maschinellen, quantitativen Analysen von Textmaterial große Erwartungen an neue Möglichkeiten für die Forschung verbunden. Neben der Bilddigitalisierung wird daher immer häufiger auch eine Erfassung des Volltextes gefordert. Diese kann entweder manuell durch Transkription 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/10.1515/9783110691597-009

QURATOR: Named Entity Recognition

- Erkennung von Personen, Orten, Organisationen in unstrukturierten Volltexten
  - https://github.com/qurator-spk/sbb_ner (Demo) (Paper)
  - BERT + unsupervised & supervised pre-training
  -



Miniſterpräſident Herriot [PER] hat geſtern morgen den Vorſitzenden der Rußland [LOC]- Kommiſſion, Senator de Monzie [PER], empfangen.
Die Botſchafterkonferenz hat geſtern morgen eine Sitzung ab–gehalten, in der ſie ſich mit laufenden Angelegenheiten beſchäftigte.
Die Handelskammer von Bordeaux [LOC] hat beſchloſſen, zu der neuen franzö-

QURATOR: Named Entity Linking

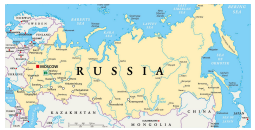
- Disambiguierung und Verlinkung von Entitäten zu Wikidata und Geokoordinaten
  - https://github.com/qurator-spk/sbb_ned (Demo) (Paper)
  - BERT Embeddings + Knowledge Base + Random Forest

Miniſterpräſident Herriot [PER] hat
geſtern morgen den Vorſitzenden
der Rußland [LOC] Kommiſſion, Senator
de Monzie [PER], empfangen.
Die Botſchafterkonferenz
hat geſtern morgen eine Sitzung ab–
gehalten, in der ſie ſich mit laufen-
den Angelegenheiten beſchäftigte.
Die Handelskammer von Bordeaux [LOC]
hat beſchloſſen, zu der neuen franzö-

?

Édouard_Herriot
Q274344 (0.75)

Russland
Q159 (0.27)



Bordeaux
Q1479 (0.58)

QURATOR: Annotation

- NER/EL Annotation (+ Transkription) direkt im Browser
  - https://github.com/qurator-spk/neat
  - Javascript + IIIF + TSV

enlarge | full

neat: neat annotation tool

User Guide | Annotation Guidelines | Issues

<<	LOCATION	POSITION	TOKEN	NE-TAG	NE-EMB	ID >>
	9	10	wäre	O	O	-
	10	11	,	O	O	-
	11	12	wenn	O	O	-
	12	13	nicht	O	O	-
	13	14	Herr	O	O	-
	14	15	Gambetta	B-PER	O	Q295090
	15	16	als	O	O	-
	16	17	deus	O	O	-
	17	18	ex	O	O	-
	18	19	machina	O	O	-
	19	20	erſchienen	O	O	-
	20	21	wäre	O	O	-

TEXT >>

6. O ſit ſich nun geſumbler gehan / iu
Fuß biß in diẽ tauſent Mann / die hatten
Tǒwens muhte / ſie kament hin gehn But-
tisholtz / da fandens mengen Engler ſtoltz /
den ſie legten ins Blutte. Es war zu mahl
ein harter Streit / das kein theil woite wei-
ch. n/ ſie ſtunden veſt zu beider ſeit / letſtlich
fleng an zaweichen / der Englich kauff vnd
nahm die flucht / alſo handt die Eydgnoſ-
ſen/ ihnen ſelber gemachet lufft.

OK CANCEL

11
12
13
14

trobert ein ſchǒne beuih / an geſchmeidt
harniſch vnd Roſſen / zwey hundert Man
hands erſchlagen / die warhrit thun ich
luch ſagen / von den freirn Eydgnoſſen /

ſ ꝛ Æ Œ æ œ å ǒ ů Å Ǒ Ů □ backspace

0 1 2 3 4 5 6 7 8 9

⸗ — ‹ › » « „ ” ’ £ § †

Mensch.Maschine.Kultur

- Neues KI-Projekt dank Förderung durch die Beauftragte des Bundes für Kultur und Medien für einen Zeitraum von 36 Monaten
- Mix aus 50% Forschung und Entwicklung und 50% Implementierung für konkrete Anwendungsszenarien und Dienste
- Aufgrund von Umfang und Vielfalt der Aufgaben ist das Projekt in vier Teilprojekte mit insgesamt 10 Arbeitspaketen unterteilt

- https://blog.sbb.berlin/mensch-maschine-kultur-neues-projekt-zur-kuenstlichen-intelligenz/

Mensch.Maschine.Kultur: TP1

- Intelligente Verfahren für die generische Dokumentanalyse
  - AP1.1: Multimodale Layouterkennung
    - Weiterentwicklung der Layoutanalyse unter Einbeziehung mehrerer Modalitäten (Bild, Geometrie, Text) und Architekturen (Vision Transformer, Graph Neural Networks)
    - Erweiterung der erkennbaren Regionentypen (bspw. Werbung, Fußnoten, Captions)
    - Erkennung von Seitenübergreifenden Strukturen (bspw. Verweise, Inhaltsverzeichnisse)
  - AP1.2: Multimodale Informationsextraktion
    - Weiterentwicklung von NER/NEL unter Einbeziehung mehrerer Modalitäten (Layout)
    - Erweiterung der erkennbaren Klassen (bspw. Autor, Herausgeber, Verlag)
  - AP1.3: Transformer für die Texterkennung
    - Bessere, schneller und robustere Methoden und Modelle für OCR mit Transformern
  - AP1.4: Texterkennung für Asiatica
    - Erweiterung der Layoutanalyse für “left-to-right” und vertikalen Text
    - Texterkennung für asiatische Schriften (Chinesisch, Japanisch, Koreanisch)

Mensch.Maschine.Kultur: TP2

- Bildanalyseinstrumente zur Erschließung des digitalen Kulturellen Erbes
  - AP2.1: Weiterentwicklung der Bildsuche
    - Erweiterung und Verbesserung der Bildextraktion, Segmentierung und Klassifikation durch Saliency, lokale Merkmale und Fine-Tuning von Modellen
    - Automatisierte Generierung von Bildbeschreibungen (Captions) für die Textbasierte Suche
    - Annotationsumgebung für die Erstellung von Trainingsdaten auf Basis von IIIF
    - Erweiterung auf zusätzliche Datenquellen wie Zeitungen, BPK
  - AP2.2: Bildsuche für Fachanwendungen
    - Aufbau von mindestens 3 fachspezifischen Anwendungen für die Bildsuche
      - Stempel und Siegel
      - Wasserzeichen
      - Druckermarken

Mensch.Maschine.Kultur: TP3

- KI-unterstützte Inhaltsanalyse und Sacherschließung
  - AP3.1: Semi-automatisierte KI-Verfahren für die Sacherschließung
    - Evaluation von Werkzeugen für die KI-unterstützte Sacherschließung (z.B. Annif)
    - Erstellung von Trainingsdaten basierend auf Katalogdaten und Klassifikationssystemen z.B. Gemeinsame Normdatei (GND), Standardthesaurus Wirtschaft (STW), Regensburger Verbundklassifikation (RVK), Basisklassifikation (BK), Klassifikation der Fachdatenbank "Index Theologicus" (IxTheo-Klassifikation)
    - Übertragung der Daten in das SKOS-Format
    - Einbringen der Ergebnisse in die Weiterentwicklung des Digitalen Assistenten DA-3
  - AP3.2: Voll-automatisierte KI-Verfahren für die Discovery
    - Integration von NER/NEL in die Suche und Indizierungsprozesse der Digitalisierten Sammlungen sowie Discovery-Umgebung
    - Visualisierungen, Suche und Recherche nach Named Entities
    - Digital Humanities use cases (z.B. historische Soziale Netzwerkanalyse)

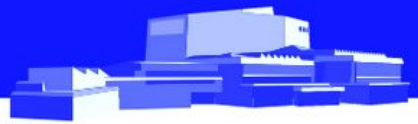
Mensch.Maschine.Kultur: TP4

- Datenbereitstellung und Kuratierung für KI
  - AP4.1: Aufbau und Vermittlung von Kompetenzen zur Erstellung und Veröffentlichung von Kulturdatensets
    - Erhebung der Anforderungen für die Veröffentlichung von digitalen Kulturdaten für KI
    - Identifizierung geeigneter Formate und Plattformen für Datenveröffentlichungen
    - Richtlinien für die Kuratierung und Veröffentlichung von digitalen Kulturdaten unter ethischen, rechtlichen und sozialen Aspekten
  - AP4.2: Aufbau von technischen Verfahren und Prozessen für die Erstellung und Veröffentlichung von digitalen Kulturdaten als Datensets
    - Entwicklung von Prozessen für die dynamische Zusammenstellung von Datensets anhand formaler, technischer und inhaltlicher Kriterien
    - Implementierung von Verfahren zur automatisierten Konvertierung von digitalen Kulturdaten in für die KI geeignete Datenformate
    - Veröffentlichung von min. 3 für KI aufbereiteten Datensets (Bilder, Texte, Metadaten)

Ethische, rechtliche und soziale Aspekte von KI

"*In practice, data collection in significant ML subfields is done without following a rigorous procedure or set of guidelines*" – Timnit Gebru, 2021

- Daten, die für das Trainieren von KI Modellen verwendet werden, haben leider zu häufig Qualitäts- und/oder ethische Probleme (siehe z.B. exposing.ai, excavating.ai u.v.m.)
- Bibliotheken verfügen über viel Kuratierungserfahrung und Qualitätsstandards
- Internationale Community zu KI und GLAM, z.B. EuropeanaTech TF AI, AI4LAM, Cultural-AI, arbeitet an Kuratierungsstandards und Empfehlungen für KI Daten und Modelle, die ethische, soziale und rechtliche Aspekte berücksichtigen (z.B. "Datasheets for Digital Cultural Heritage")
- Teilprojekt 3 von "Mensch.Maschine.Kultur" wird durch eine Ethical Foresight Analysis begleitet
- Wie können Bibliotheken zu einem verantwortungsvollen Umgang mit Daten in Forschung und Anwendung von KI beitragen?
- Welche Verfahren und Praktiken in Bibliotheken müssen hinterfragt, welche Biases offengelegt werden?



Staatsbibliothek
zu Berlin
Preußischer Kulturbesitz

Danke für die Aufmerksamkeit!

Fragen?

Clemens Neudecker | Staatsbibliothek zu Berlin — Preußischer Kulturbesitz
5. KOBV-Fachkolloquium 2022 | 28. November 2022 | Zuse Institut Berlin