

**Staatsbibliothek  
zu Berlin**  
Preußischer Kulturbesitz



# **Künstliche Intelligenz für das digitale kulturelle Erbe in der Staatsbibliothek zu Berlin**

Clemens Neudecker | Staatsbibliothek zu Berlin – Preußischer Kulturbesitz  
Workshop “Einsatz von KI und DH in Bibliotheken”  
3. & 4. November 2022 | Deutsche Nationalbibliothek, Frankfurt am Main

# Datengrundlage @ SBB

- >2,5 PetaByte digitale Daten
  - Digitalisierte Sammlungen
    - <https://digital.staatsbibliothek-berlin.de/>
    - > 200,000 Werke
    - > 20 Mio. Images
  - Digitalisierte Zeitungen (ZEFYS)
    - <https://zefys.staatsbibliothek-berlin.de/>
    - > 7 Mio. Images
  - Fachinformationsdienste
    - > 200 Mio. digitale Objekte
  - SBB LAB
    - <https://lab.sbb.berlin/>
    - APIs, Datensets, Demos, Hackathons

# Projekte und Personal @ SBB

- OCR-D (DFG)
  - <https://ocr-d.de/>
  - 12/2016 – 06/2024
  - 1x 100% E13 (DevOps)
- 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)

# OCR-D

- Hauptaufgaben der SBB im OCR-D Koordinierungsprojekt
  - Entwicklung und Abstimmung von technischen Spezifikationen  
<https://ocr-d.de/en/spec/>
  - Entwicklung der OCR-D Python Referenzimplementierung `core`  
<https://github.com/OCR-D/core>
  - Release Management von `ocrd_all`  
[https://github.com/OCR-D/ocrd\\_all](https://github.com/OCR-D/ocrd_all)
  - Pflege der technischen Dokumentation  
<https://ocr-d.de/en/use>
  - Community Support in div. Kanälen und Formaten  
<https://ocr-d.de/en/community>
- Beantragt: Weiterentwicklung und Optimierung der Layoutanalyse
- KI in OCR-D: Mehrere Tools (z.B. Layoutanalyse, Texterkennung) verwenden neuronale Netze und trainierte Modelle und profitieren vom Einsatz von GPUs

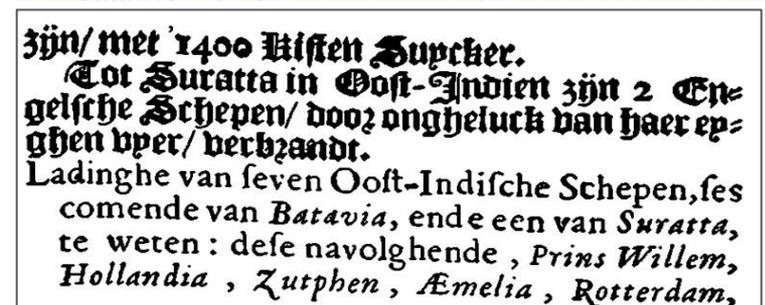
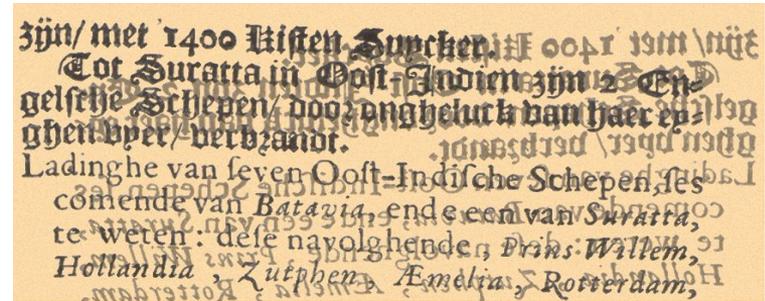
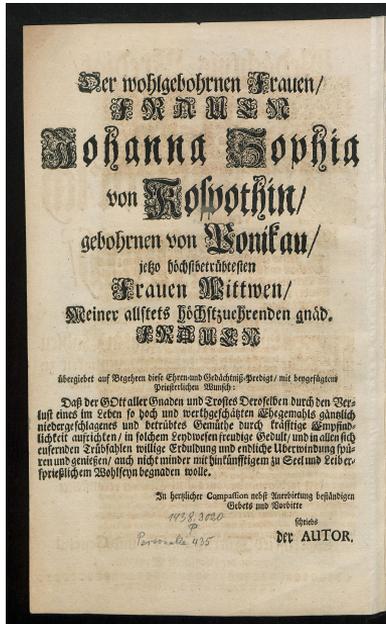
# QURATOR

- 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)



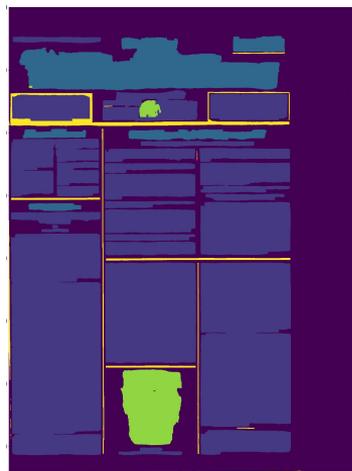
# QURATOR: Binarisierung

- Binarisierung für Texterkennung
  - [https://github.com/qurator-spk/sbb\\_binarization](https://github.com/qurator-spk/sbb_binarization)
  - Hybrid CNN + Transformer



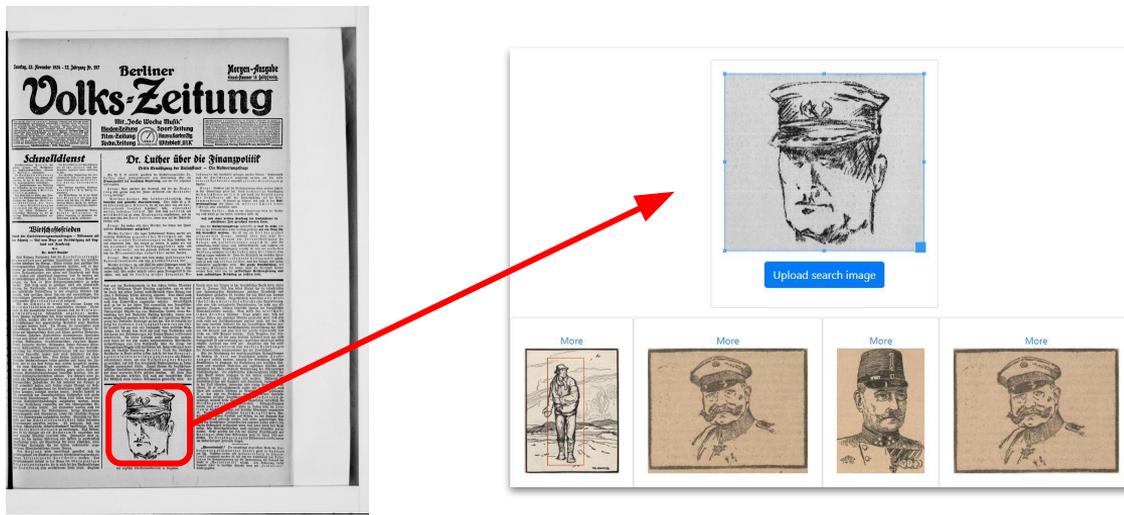
# QURATOR: Eynollah

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



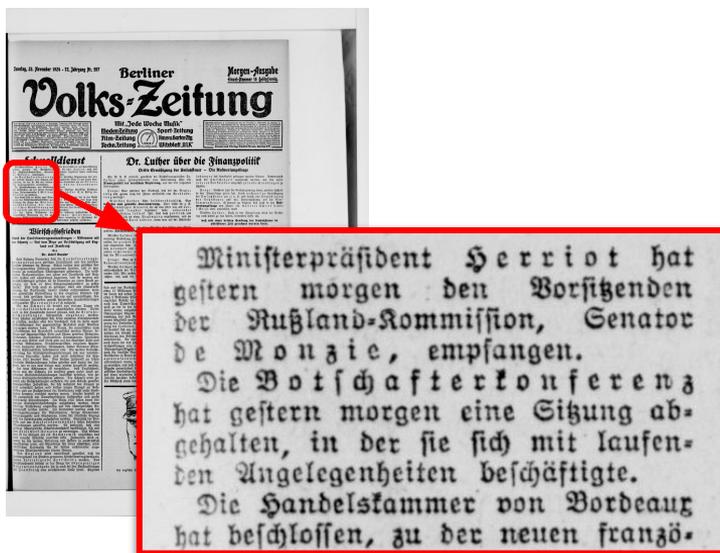
# QURATOR: Bildähnlichkeitssuche

- Bildähnlichkeitssuche und Bildklassifikation
  - [https://github.com/qurator-spk/sbb\\_images](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](https://github.com/qurator-spk/ocrd_calamari)
  - CNN + LSTM + CTC

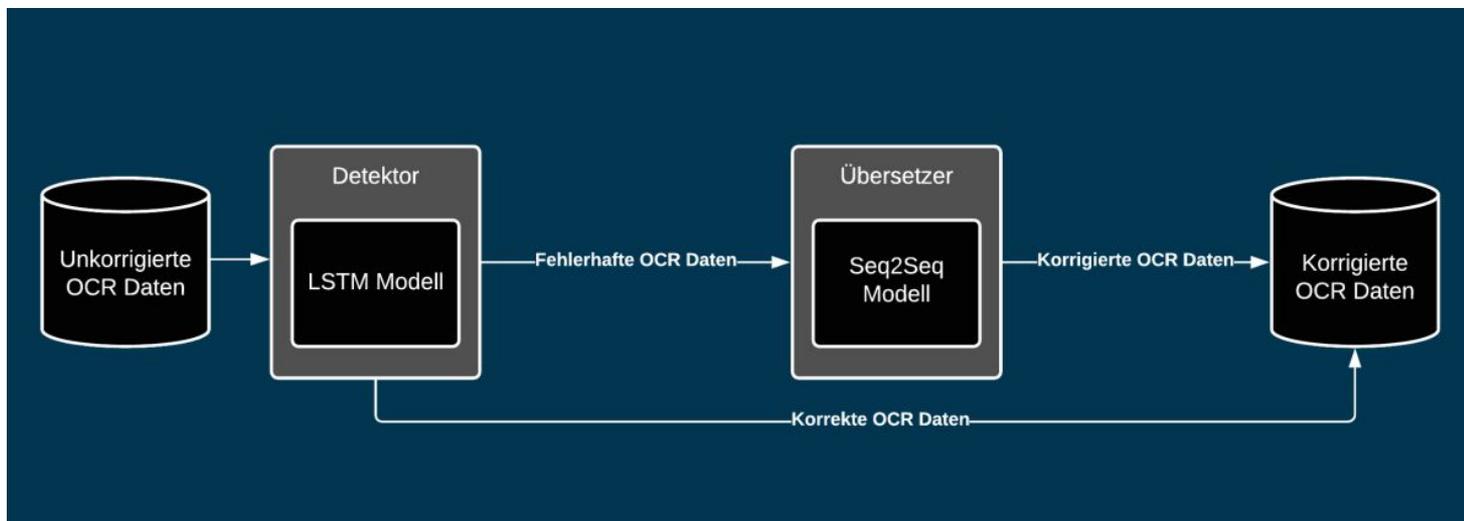


Ministerpräsident Herriot hat die Feierlichkeiten zur Ueberführung geiern morgen de» Vorsitzenden der Autlanü-Kommission, Senator d. M. O. Il z i e, empfangen. Die Botschafterkonferenz bat gestern morgen eine Sitzung abgehalten, in der sie sich mit laufenden Angelegenheiten beschäftigte. Die Handelskammer von Bordeaux bat defchloffen, zu der neuen französischen Inlandsanleihe 1 Million Francs zu zeichne». Aus Genf sind in Sofia zwei Delegierte der Bälkerbundskommission zur Prüfung der Frage der Massen, auswanderung der bulgarischen Bevölkerung aus Thrazien und Mazedonien und der letzien Beschwerde der bulgarischen Regierung an die zu ständige Dölkerbundskommission getroffen.

Ministerpräsident Herriot hat gestern morgen den Vorsitzenden der Rußland-Kommission, Senator de Monzie, empfangen. Die Botfchafterkonferenz hat gestern morgen eine Sitzung abgehalten, in der sie sich mit laufenden Angelegenheiten beschäftigte. Die Handelskammer von Bordeaux hat bechloffen, zu der neuen französischen Inlandsanleihe 1 Million Francs zu zeichnen. Aus Genf sind in Sofia zwei Delegierte der Völkerbundskommission zur Prüfung der Frage der Massenwanderung der bulgarischen Bevölkerung aus Thrazien und Mazedonien und der letzten Befchwerde der bulgarischen Regierung an die zu ständige Völkerbundskommission getroffen.

# QURATOR: Nachkorrektur (OCR)

- Automatisierte Nachkorrektur von OCR Resultaten
  - [https://github.com/qurator-spk/sbb\\_ocr\\_postcorrection](https://github.com/qurator-spk/sbb_ocr_postcorrection) (Paper)
  - Sequence-to-Sequence LSTM





# QURATOR: Named Entity Recognition

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



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-

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-

# QURATOR: Named Entity Linking

- Disambiguierung und Verlinkung von Entitäten zu Wikidata und Geokoordinaten
  - [https://github.com/qurator-spk/sbb\\_ned](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-

?

Russland  
Q159 (0.27)



Édouard\_Herriot  
Q274344 (0.75)

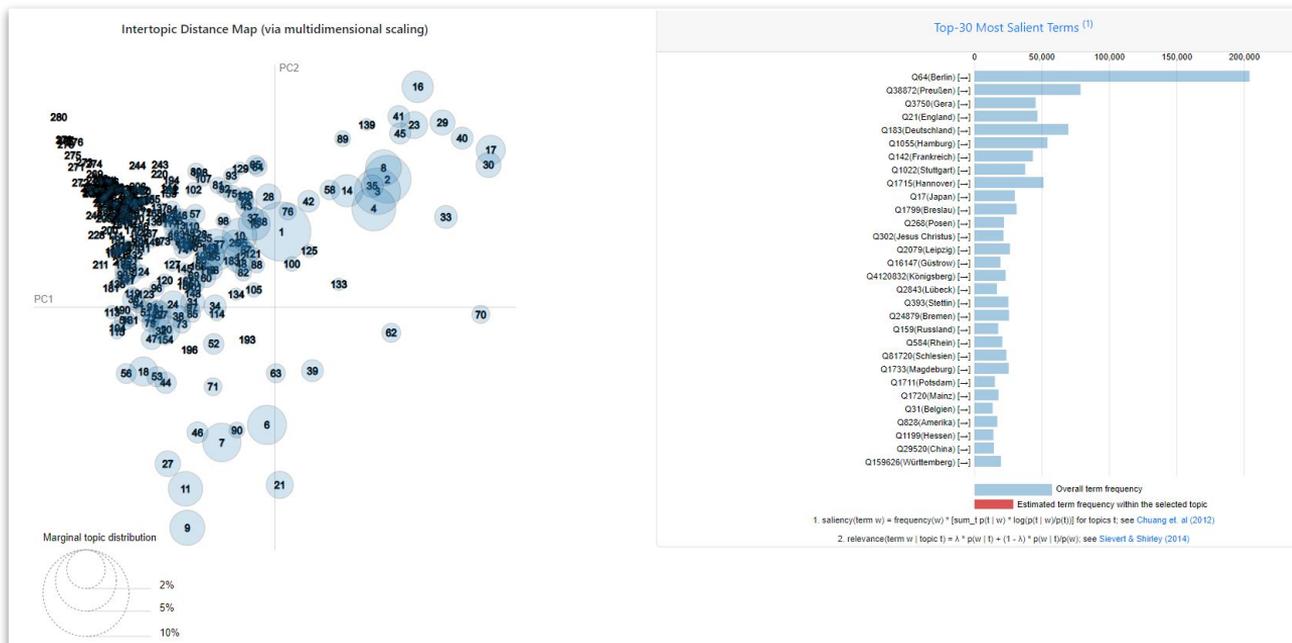


Bordeaux  
Q1479 (0.58)



# QURATOR: Topic Modelling

- Topic Modelling
  - [https://github.com/qurator-spk/sbb\\_topic-modelling](https://github.com/qurator-spk/sbb_topic-modelling) (Demo)
  - Gensim (LDA)



# QURATOR: Annotation

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

neat: neat annotation tool

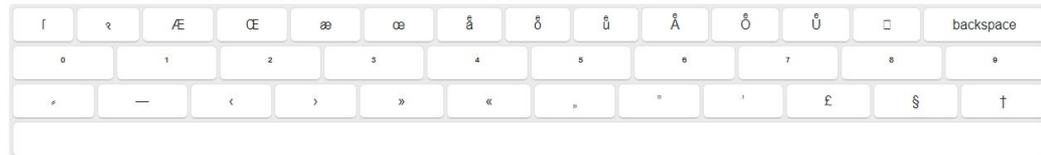
[User Guide](#) | [Annotation Guidelines](#) | [Issues](#)



[enlarge](#) | [full](#)

<<	LOCATION	POSITION	TOKEN	NE-TAG	NE-EMB	ID >>	TEXT >>
	9	10	wäre	O	O	-	6. O fit sich nun gefumbler gehan / iu
	10	11	,	O	O	-	Fuß biß in dië taufent Mann / die hatten
	11	12	wenn	O	O	-	Töwens muhte / fie kament hin gehn But-
	12	13	nicht	O	O	-	tisholtz / da fandens mengen Engler ftoltz /
	13	14	Herr	O	O	-	den fie legten ins Blutte. Es war zu mahl
	14	15	Gambetta	B-PER	O	Q295090	ein harter Streit / das kein theil wolte wei-
	15	16	als	O	O	-	ch. n/ fie ftunden velt zu beider feit / lefftlich
	16	17	deus	O	O	-	fleng an zaweichen / der English kauff vnd
	17	18	ex	O	O	-	nahm die flucht / alfo handt die Eydgnof-
	18	19	machina	O	O	-	fen/ ihnen felber gemachet lufft.
	19	20	erfchienen	O	O	-	
	20	21	wäre	O	O	-	

11 trobert ein fchöne beuh / an gelchmeidt  
12 harnifch vnd Roffen / zwey hundert Man  
13 hands erfchlagen / die warhrit thun ich  
14 luch fagen / von den freim Eydgnoffen /



# Mensch.Maschine.Kultur

- Förderung durch die Beauftragte des Bundes für Kultur und Medien für einen Zeitraum von 36 Monaten (KI Strategie des Bundes)
- Mix aus 50% Forschung und Entwicklung mit 50% Implementierung für konkrete Anwendungsszenarien und Dienste
- Aufgrund des Umfangs und der Vielfalt der Aufgaben ist das Projekt in wiederum 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 “right-to-left” und vertikalen Text
    - Texterkennung für asiatische Schriften (Chinesisch, Japanisch, Koreanisch)

# Mensch.Maschine.Kultur: TP2

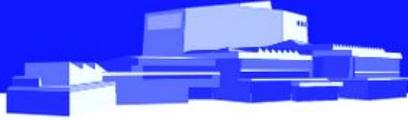
- Bildanalyseinstrumente zur Erschließung des digitalen Kulturellen Erbes
  - AP2.1: Weiterentwicklung der Bildsuche
    - Erweiterung und Verbesserung der Bildextraktion, Segmentierung und Klassifikation durch Saliency, lokale Merkmale und Fine-Tuning von Modellen
    - Automatisierte Generierung von 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)



**Staatsbibliothek  
zu Berlin**  
Preußischer Kulturbesitz



**Danke für die Aufmerksamkeit!**

**Fragen?**

Clemens Neudecker | Staatsbibliothek zu Berlin – Preußischer Kulturbesitz  
Workshop “Einsatz von KI und DH in Bibliotheken”  
3. & 4. November 2022 | Deutsche Nationalbibliothek, Frankfurt am Main