Digitalisate kuratieren mit KI - von unstrukturierten Daten zu strukturierten Inhalten

Clemens Neudecker (@cneudecker)
Staatsbibliothek zu Berlin – Preußischer Kulturbesitz

Fachtagung Netzwerk maschinelle Verfahren in der Erschließung
10.-11. Oktober 2019, DNB, Frankfurt am Main

Staatsbibliothek zu Berlin
Preußischer Kulturbesitz

Qurator
Curation Technologies

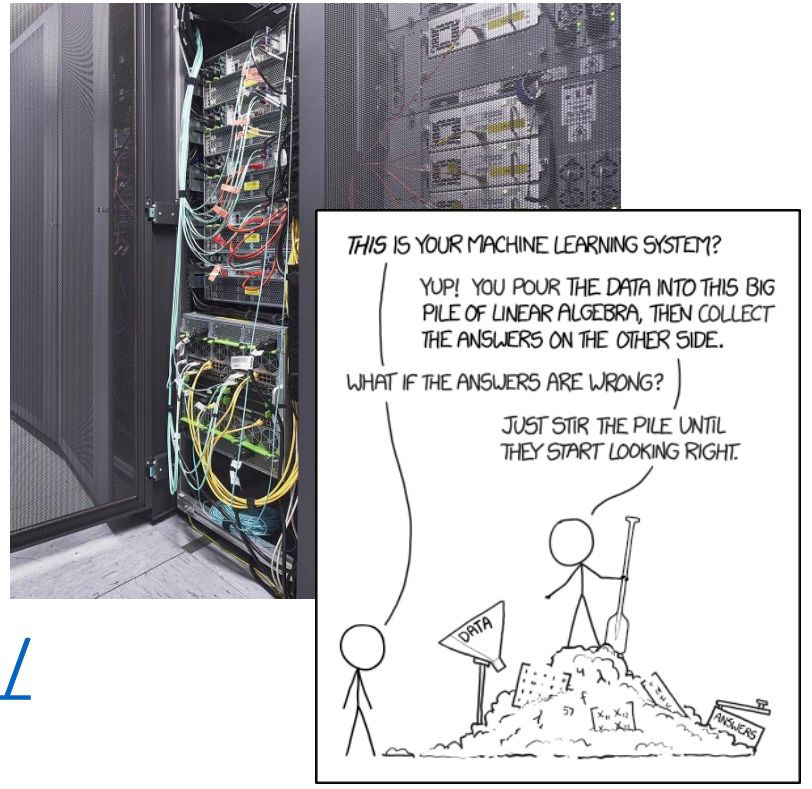
Hintergrund

- > 12 Mio. Dokumente
- Metadaten (METS, MODS)
- Digitalisierte Sammlungen
  - https://digital.staatsbibliothek-berlin.de/
  - ca. 160,000 Digitalisate
  - ca. 5 Mio. Seiten OCR
- Digitalisierte Zeitungen
  - http://zefys.staatsbibliothek-berlin.de/
  - ca. 7 Mio. Seiten digitalisiert
  - ca. 3 Mio. Seiten OCR
- ca. 2,5 PetaBytes Daten

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
WACHSTUMSKERNE
UNTERNEHMEN REGION
Die BMBF-Innovationsinitiative Neue Länder

QURATOR @ SBB

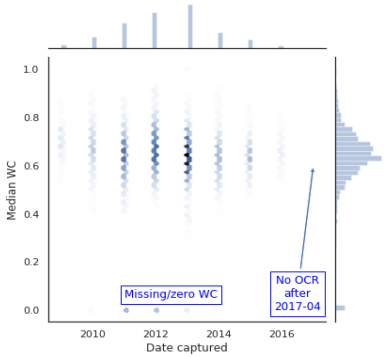
- Gemeinsame Projektsteuerung durch Forschungsreferent und Referatsleiter IT-Anwendungen und Datenmanagement
- 3x FTE Entwicklerstellen E13 FuE (36 Monate)
- SPK-KI Server:
  - 2x Nvidia Tesla V100 24GB (VGPU)
  - 36x Intel XEON 2.7 Ghz
  - 192GB RAM
- Freie Bereitstellung von SBB Daten, Technologien und Anwendungen: https://github.com/qurator-spk
- Mehr zur SBB in QURATOR: https://qurator.ai/partner/staatsbibliothek-zu-berlin/
- SBB Blogserie „Künstliche Intelligenz“: https://blog.sbb.berlin/tag/wissenschaftsjahr-2019/



https://xkcd.com/1838/

Metadatenanalyse

- Analyse von Metadaten (METS/MODS) und bestehenden Volltexten (ALTO) zur Gewinnung von Informationen über Merkmale und deren Verteilung



WC-Analyse

```
Mean squared error: 0.0040
R² explained variance score: 0.50
```

	y_test	y_pred
PPN736019553	0.440000	0.448807
PPN622773135	0.693333	0.727628
PPN821772082	0.850000	0.673339
PPN717743837	0.472500	0.525989
PPN777540037	0.627500	0.650604
PPN666470626	0.665000	0.696438
PPN679959122	0.751111	0.661616
PPN735327963	0.557500	0.527351
PPN655207724	0.555556	0.637708
PPN816552711	0.521786	0.520847
PPN799319708	0.616667	0.617835
PPN755765214	0.720000	0.656037
PPN749791160	0.628333	0.692473
PPN746510098	0.553333	0.576473
PPN728337703	0.757143	0.705831

```
0.5630313527894495 originInfo-publication0_dateIssued
0.06318379892891454 classification-ZVDD_Ostasiatica
0.03823449159666124 originInfo-publication0_place_placeTerm_London
0.029608480124117924 language_languageTerm_ger
0.029425594614164 41 originInfo-digitization0_dateCaptured
0.028823039309747406 subject-EC1418_genre_book
0.028199724369852127 classification-ZVDD_Rechtswissenschaft
0.023531983883285302 originInfo-publication0_place_placeTerm_Emden
0.023497157041718043 genre-aad_Liedersammlung
0.022995013108022253 originInfo-publication0_publisher_Clark
0.022987499197938 87 originInfo-publication0_place_placeTerm_Neudamm-Berlin
0.01570735000960244 file_count
0.013657140133976932 word_count
0.010590282448155778 classification-ZVDD_Kunst
0.009108230229550436 subject-EC1418_genre_journal
0.008340169667697 41 classification-ZVDD_Musik
0.006893187886055325
0.005942084560844258
0.005686767728023 92
0.004100967591525305
```

Feature Ranking

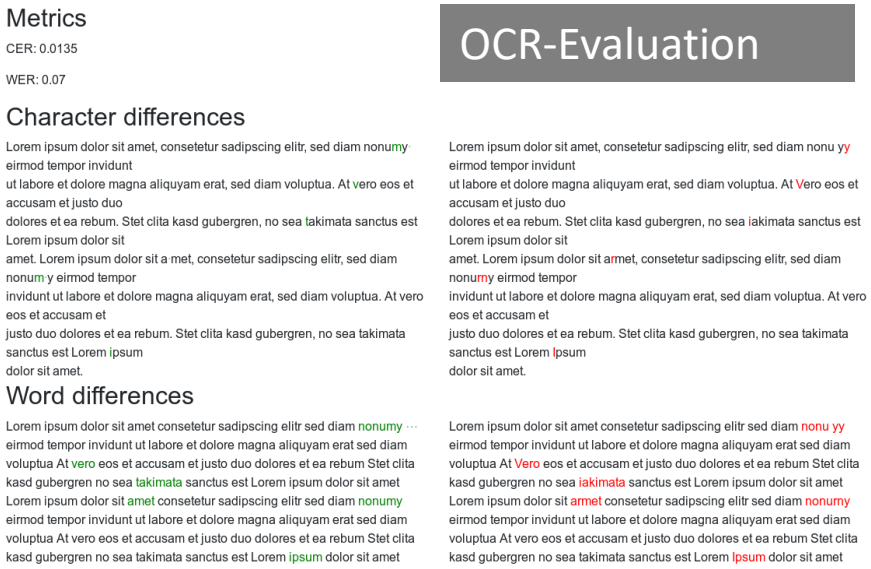
https://github.com/qurator-spk/modstool

	message
Changed w3cdtf encoding to iso8601	101757
Fixed eventType for a created origin	43254
Fixed eventType for electronic ed.	38851
Forced single instance of {http://www.loc.gov/mods/v3}dateCreated	36854
Fixed eventType for an issued origin	14300
Not a iso8601 date: "o.D."	2466
Forced single instance of {http://www.loc.gov/mods/v3}dateIssued	2105
Filtered {http://www.loc.gov/mods/v3}originInfo element (has no eventType)	1361
Forced single instance of {http://www.loc.gov/mods/v3}subject	762
Changed scriptTerm authority to lower case	437
Not a iso8601 date: "[18. Jh.]"	142
Not a iso8601 date: "1800 (1800c)"	124
Not a iso8601 date: "1780 (1780c)"	115
Not a iso8601 date: "[19. Jh.]"	86
Not a iso8601 date: "[15. Jh.]"	67

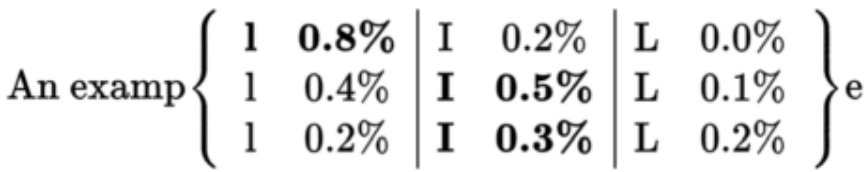
Metadaten-Validierung

OCR Evaluation & Qualitätsverbesserung

- OCR Evaluation und Qualitätsverbesserung digitalisierter Dokumente durch bessere OCR und automatisierte OCR-Nachkorrektur



OCR-Evaluation



Varianten-Voting in Calamari OCR

- https://github.com/qurator-spk/dinglehopper
- https://github.com/qurator-spk/ocrd_calamari (trainiert auf GT4HistOCR)

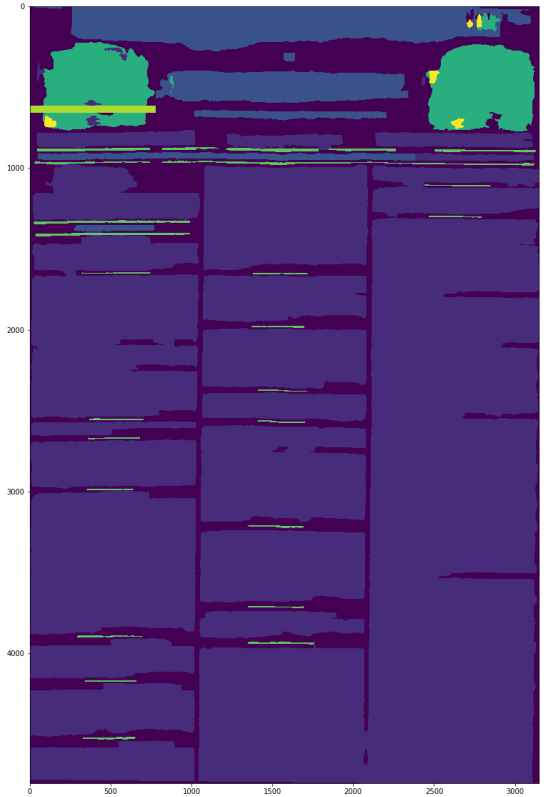
OCR Exkurs: OCR-D

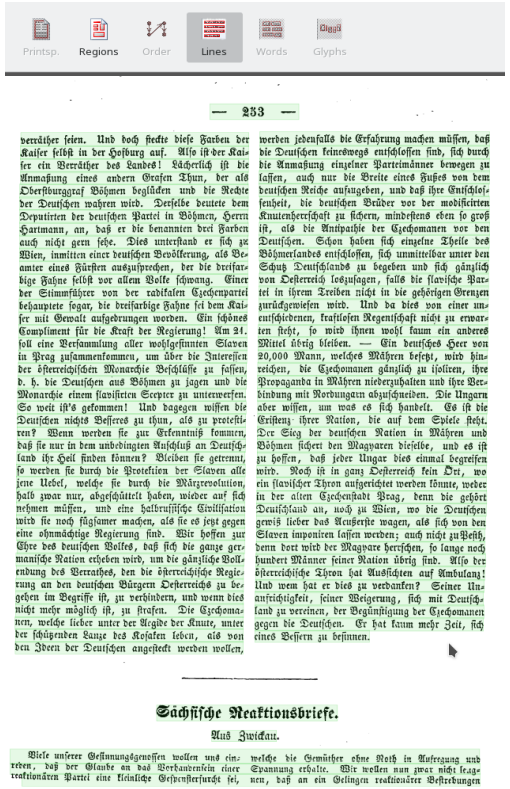
- DFG-gefördertes Koordinierungsprojekt + 8 verteilte Forschungsprojekte zu OCR für historische Drucke

- Webseite: http://ocr-d.de/
- Module: https://ocr-d.github.io/projects
- Source Code: https://github.com/OCR-D
- Dokumentation: https://ocr-d.github.io/
- Ground Truth: http://ocr-d.github.io/gt-repo
- Chat: https://gitter.im/OCR-D/Lobby

Layout- bzw. Strukturerkennung

- Erkennung und Klassifikation von Strukturmerkmalen: https://github.com/qurator-spk/pixelwise_segmentation_SBB

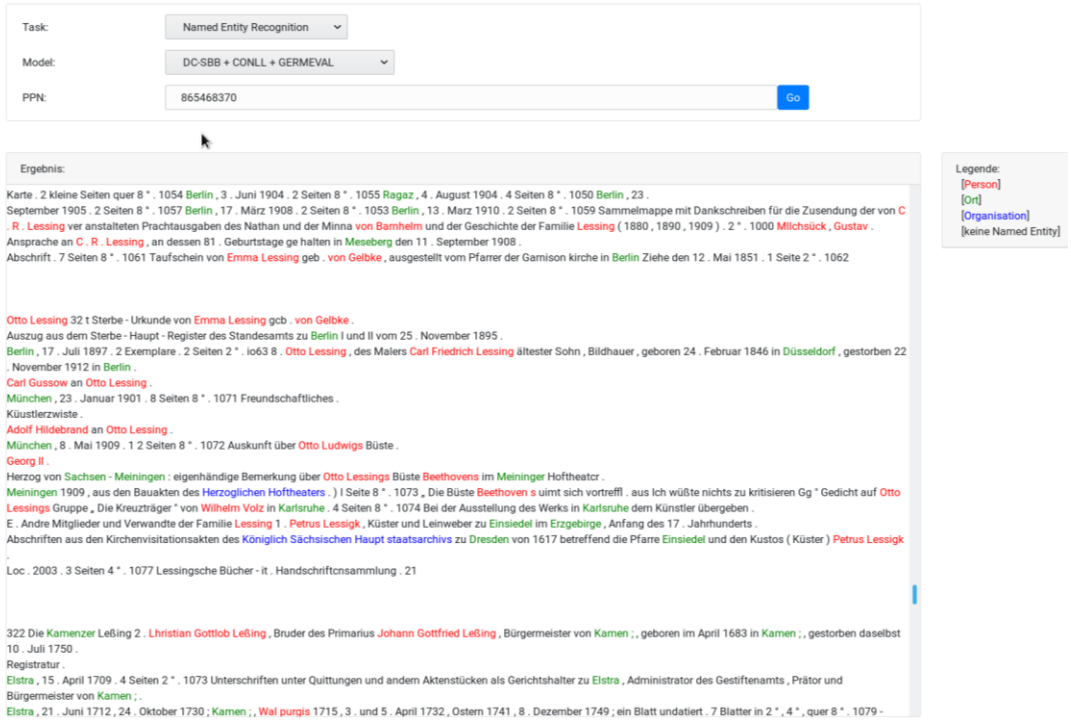




- Pixel-Labelling mit ResNet50/UNet (CNN) für aktuell 16 Objektklassen
  - Spalten, Absätze, Separatoren
  - Überschriften, Fußnoten, Marginalien
  - Tabellen, Grafiken
  - usw.
- Textzeilenextraktion für die OCR
- Erkennung der Lese- bzw. Artikelreihenfolge (Reading Order)

Named Entity Recognition

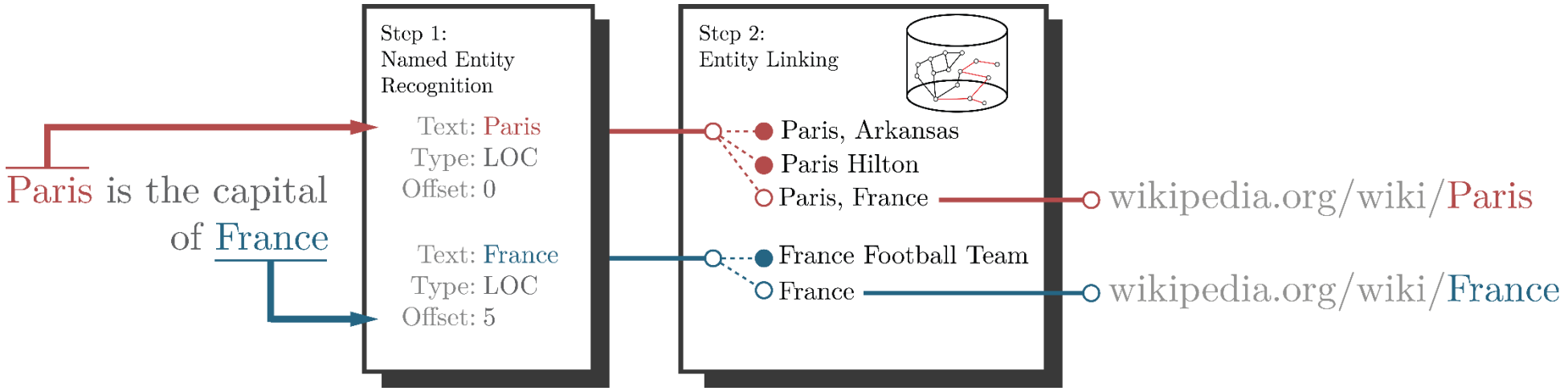
- Erkennung und Klassifikation benannter Entitäten in digitalisierten Dokumenten mit `BERT`: https://github.com/qurator-spk/sbb_ner



		BERT multi-lingual-cased			(Riedl and Padó, 2018)	(Schweter and Baiter, 2019)
5-fold cross validation on	pre-train	precision	recall	F_1	F_1	F_1
SBB	DC-SBB + GermEval + CoNLL	81.1 ±1.2	87.8 ±1.4	84.3 ±1.1	-	-
	DC-SBB + CoNLL	81.0 ±2.1	87.6 ±1.8	84.2 ±1.9	-	-
	DC-SBB + GermEval	80.6 ±1.8	87.4 ±1.3	83.8 ±1.2	-	-
	CoNLL	81.0 ±1.9	86.6 ±2.2	83.7 ±1.5	-	-
	GermEval	79.7 ±1.8	87.2 ±0.8	83.3 ±1.1	-	-
	GermEval + CoNLL	79.9 ±2.1	86.4 ±1.7	83.0 ±1.9	-	-
	DC-SBB	79.1 ±2.6	86.7 ±0.7	82.7 ±1.3	-	-
	none	79.1 ±3.6	85.0 ±1.1	81.9 ±2.2	-	-
ONB	Newspaper (1703-1875)	-	-	-	-	85.31
	DC-SBB+GermEval + CoNLL	81.5 ±1.8	87.8 ±1.4	84.6 ±1.5	-	-
	DC-SBB + GermEval	81.6 ±2.5	87.5 ±1.6	84.5 ±1.8	-	-
	DC-SBB + CoNLL	81.7 ±2.8	87.5 ±1.9	84.5 ±2.3	-	-
	DC-SBB	81.8 ±2.3	87.1 ±2.1	84.3 ±2.0	-	-
	GermEval	80.8 ±2.1	85.4 ±1.2	83.0 ±1.4	78.56	-
	GermEval + CoNLL	80.0 ±1.5	84.7 ±1.6	82.3 ±1.5	-	-
	CoNLL	79.1 ±2.5	84.5 ±2.1	81.7 ±2.2	76.17	-
	none	78.0 ±2.4	84.1 ±1.9	80.9 ±2.0	73.31	-
LFT	Newspaper (1888-1945)	-	-	-	-	77.51
	DC-SBB + CoNLL	70.0 ±2.6	81.0 ±0.7	75.1 ±1.5	-	-
	DC-SBB + GermEval	69.9 ±3.0	81.1 ±1.0	75.1 ±1.8	-	-
	DC-SBB	70.0 ±3.5	80.8 ±1.4	75.0 ±2.1	-	-
	DC-SBB + GermEval + CoNLL	69.8 ±3.0	80.8 ±0.9	74.9 ±2.0	-	-
	GermEval	68.9 ±2.7	79.3 ±1.4	73.7 ±1.9	74.33	-
	GermEval + CoNLL	69.1 ±2.6	78.8 ±1.3	73.6 ±1.5	-	-
	none	68.8 ±3.4	79.2 ±1.5	73.6 ±2.2	69.62	-
	CoNLL	68.4 ±3.1	79.1 ±1.3	73.3 ±2.1	72.9	-

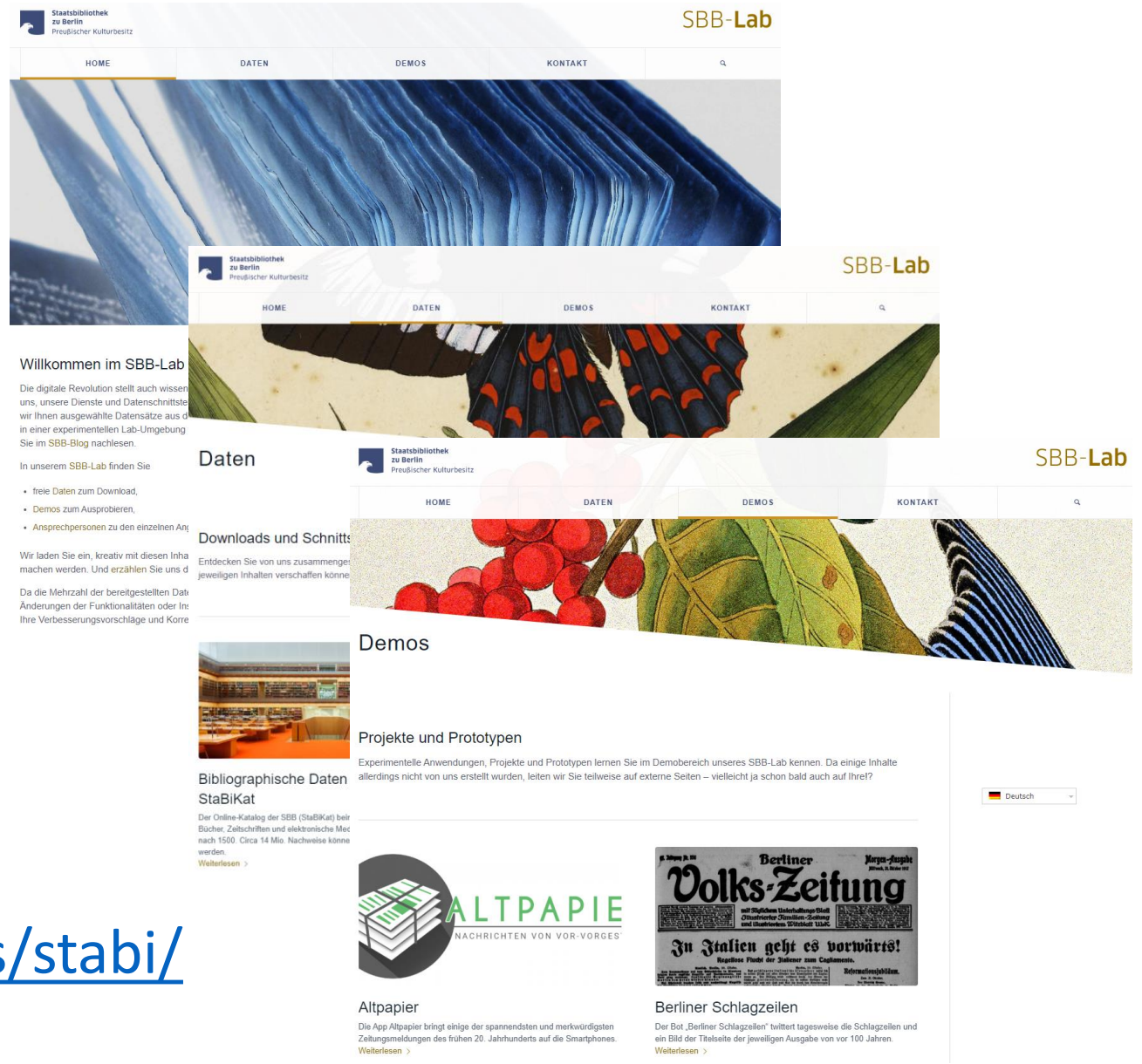
Named Entity Disambiguation & Linking

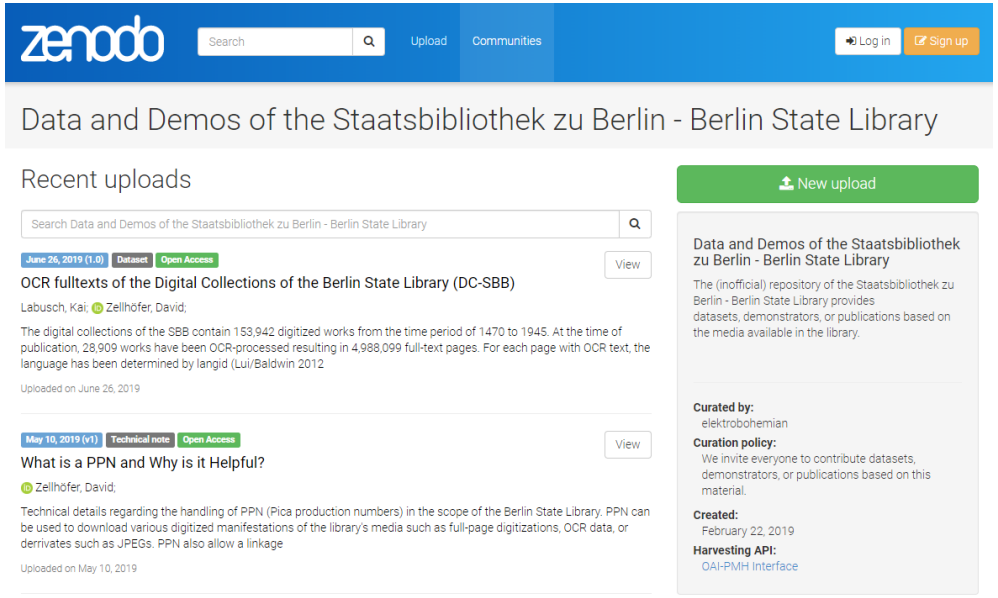
- Disambiguierung und Verlinkung benannter Entitäten mit einer Knowledge Base (Wikidata, GND)
- Erster Ansatz basierend auf Embeddings (Fasttext & Flair)



Daten & Modelle

- https://lab.sbb.berlin/

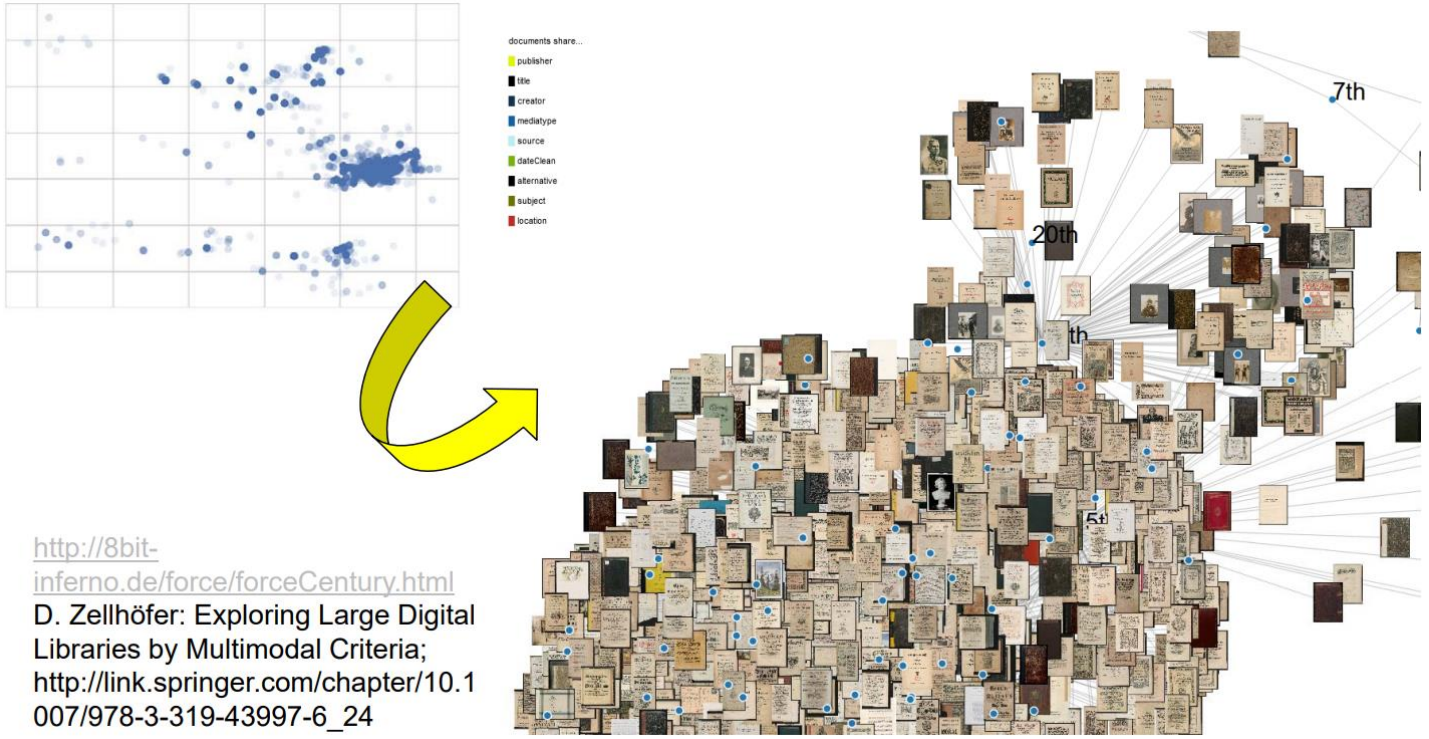




- https://zenodo.org/communities/stabi/

Ausblick

- Bildähnlichkeitssuche unter Verwendung von VGG16 und Re-training mit ImageNet
- Geolokalisierung durch Kombination von semantischen mit topographischen Merkmalen



http://8bit-inferno.de/force/forceCentury.html
D. Zellhöfer: Exploring Large Digital Libraries by Multimodal Criteria; http://link.springer.com/chapter/10.1007/978-3-319-43997-6_24

- Demonstrator basierend auf digitalisierten Sammlungen (SBB Lab)

Danke für die Aufmerksamkeit!
Fragen?

Clemens Neudecker (@cneudecker)
Staatsbibliothek zu Berlin – Preußischer Kulturbesitz
Fachtagung Netzwerk maschinelle Verfahren in der Erschließung
10.-11. Oktober 2019, DNB, Frankfurt a.M.

Staatsbibliothek zu Berlin
Preußischer Kulturbesitz

Qurator
Curation Technologies